Nonlinear Filtering in High Dimension

Patrick Rebeschini

A Dissertation

Presented to the Faculty

of Princeton University

in Candidacy for the Degree

of Doctor of Philosophy

Recommended for Acceptance

by the Department of

Operations Research and Financial Engineering

Adviser: Ramon van Handel

June 2, 2014

Abstract

The goal of filtering theory is to compute the filter distribution, that is, the conditional distribution of a stochastic model given observed data. While exact computations are rarely possible, sequential Monte Carlo algorithms known as particle filters have been successfully applied to approximate the filter distribution, providing estimates whose error is uniform in time. However, the number of Monte Carlo samples needed to approximate the filter distribution is typically exponential in the number of degrees of freedom of the model. This issue, known as curse of dimensionality, has rendered sequential Monte Carlo algorithms largely useless in high-dimensional applications such as multi-target tracking, weather prediction, and oceanography. While over the past twenty years many heuristics have been suggested to run particle filters in high dimension, no principled approach has ever been proposed to address the core of the problem.

In this thesis we develop a novel framework to investigate high-dimensional filtering models and to design algorithms that can avoid the curse of dimensionality. Using concepts and tools from statistical mechanics, we show that the decay of correlations property of high-dimensional models can be exploited by implementing localization procedures on ordinary particle filters that can lead to estimates whose approximation error is uniform both in time and in the model dimension.

Ergodic and spatial mixing properties of conditional distributions play a crucial role in the design of filtering algorithms, and they are of independent interest in probability theory. To better capture ergodicity quantitatively, we develop new comparison theorems to establish dimension-free bounds on high-dimensional probability measures in terms of their local conditional distributions. At a qualitative level, we investigate previously unknown phenomena that can only arise from conditioning in infinite dimension. In particular, we exhibit the first known example of a model where ergodicity of the filter undergoes a phase transition in the signal-to-noise ratio.

Acknowledgements

Gratitude. Profound and sincere gratitude is what moves each and every single word in this thesis.

First and foremost, gratitude towards my advisor: Ramon van Handel. Ramon's genuine passion for probability theory is what moved and inspired my Ph.D. studies since the very beginning. His deep enthusiasm and extraordinary intellectual curiosity made my experience at Princeton an amazing journey. Interacting with Ramon always felt like landing on Hyperuranion, a world where vivid ideas float around and are as easy to grasp as hand-picking apples. I would like to thank Ramon for several things. He gave me intuition. He taught me how to do research. He revealed to me what I can achieve through devotion and perseverance. He introduced me to the ideas that led to the work in this thesis while providing formidable guidance throughout. He showed me the meanings of the words "advisor" and "mentor." Ultimately, I thank Ramon for exposing how fun research can be and being a powerful example on how enthusiasm and passion can drive a career. Thank you.

Gratitude towards my teacher: Erhan Çınlar. Erhan taught me grace, beauty, and simplicity. He taught me what it means to think probabilistically. He taught me, really, what it means to think. I will forever be thankful for the countless insights he shared with me—his sharp remarks and opinions proved to me that being a scholar means to go beyond academic boundaries. Erhan showed me what a "doctor" really is: from the Latin "docere," which means "to teach." Not only did he demonstrate to me what defines excellence in teaching, he also set an extraordinary model that I can aspire to. "You can not be a great teacher unless you think highly of your students;" this and many others. I will never forget. Thank you.

Gratitude towards my family: Luciano, Ivana, Matteo, and Vanessa. They have been such a constant presence throughout my doctoral studies by giving me the deep roots to stay down to earth and judge what is truly important in life. It takes a lot of strength to let a son or a brother follow his path without holding him back. Thank you.

Gratitude towards the many friends that life has surrounded me with, luckily too many to all be mentioned here. During these years they taught me once again what "friendship" means, which is to care about someone's destiny without expecting anything in return. They kept offering their genuine friendship regardless of my ability to give back what they truly deserved. Pursuing my Ph.D. has set huge constraints on my time and location, yet I learned that true friendship goes beyond time and space. Thank you all.

Finally, since this thesis is in probability theory, gratitude goes to Tyche, the goddess of luck and randomness. For all of the above. Luckily, I am aware that many circumstances have been, still are, and always will be out of my control. She drew a wonderful realization, indeed.

To my family,
whose influence on this work
runs deeper than I can know

Contents

Chapter 1

Introduction

1.1 Nonlinear filtering and particle filters

A fundamental problem in a broad range of applications is the combination of observed data and dynamical models. Particularly in highly complex systems with partial observations, the effective extraction and utilization of the information contained in observed data can only be accomplished by exploiting the availability of accurate predictive models of the underlying dynamical phenomena of interest. Such problems arise in applications that range from classical tracking problems in navigation and robotics to extremely large-scale problems such as weather forecasting. In the latter setting, and in other complex applications in the geophysical, atmospheric and ocean sciences, incorporating observed data into dynamical models is called *data assimilation*.

From a statistical perspective, it is in principle simple to formulate the optimal solution to the data assimilation problem. We model the dynamical process that is not directly observable as a time-homogeneous Markov chain $(X_n)_{n\geq 0}$ on a measurable space $(E, \mathcal{E})$, with $\mathbf{P}(X_n \in dz | X_{n-1} = x) = p(x, z)\psi(dz)$, for a certain transition density p and reference measure ψ. We model the noisy observations $(Y_n)_{n\geq 0}$ as a collection of random variables on a measurable space $(F, \mathcal{F})$ that are conditionally independent given $(X_n)_{n\geq 0}$, with $\mathbf{P}(Y_n \in dz | X_n = x) = g(x, z)\varphi(dz)$, for a certain observation density g and reference measure φ. The joint process $(X_n, Y_n)_{n\geq 0}$ that takes values in $(E \times F, \mathcal{E} \otimes \mathcal{F})$ is called *hidden Markov model*, and its dependency structure is illustrated in Figure 1.1.

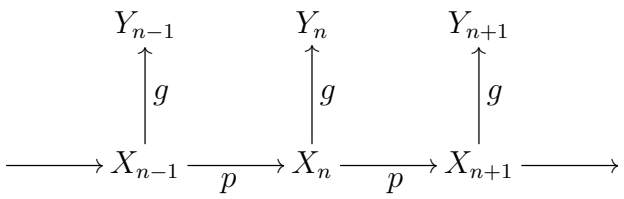

Figure 1.1: Dependency graph of a hidden Markov model.

In many applications one is interested in estimating the hidden state X_n based on the observation history $Y_1, \ldots, Y_n$ to date, and to compute $\mathbf{E}(f(X_n)|Y_1, \ldots, Y_n)$ for a certain function f, or for a certain class of functions. For instance, we might be interested in tracking the position of a boat given the noisy measurements coming from a radar, and we might want to know how accurate our estimates are. Or we might be interested in evaluating the temperature field of the weather over a certain geographical location given the noisy measurements coming from weather stations, together with the uncertainty in our estimates. More generally, one is often interested in computing the conditional mean and variance of the underlying process given the observations history.

However, in almost all cases the conditional estimates for individual functions do not form a closed system of equations, and one has to compute the *nonlinear filter* distribution

$$\pi_n := \mathbf{P}(X_n \in \cdot\,|Y_1, \ldots, Y_n).$$

If the filter π_n can be computed, it yields an optimal (in the least mean square sense) estimate of X_n given the observations up to time n, as well as a complete representation of the uncertainty in this estimate.

An important property of the filter is its recursive structure: π_n depends only on π_{n-1} and the new observation Y_n. In fact, it is easily verified using the Bayes formula (cf. Section 3.3) that π_n can be computed recursively in two steps, the so-called *prediction* and *correction* step:

$$\pi_{n-1} \xrightarrow{\text{prediction}} \mathsf{P}\pi_{n-1} \xrightarrow{\text{correction}} \pi_n = \mathsf{C}_n\mathsf{P}\pi_{n-1},$$

where P and C_n are, respectively, the prediction and correction operators that are defined as

$$(\mathsf{P}\rho)f := \int \rho(dx)\, p(x,x')\, \psi(dx')\, f(x'),$$
$$(\mathsf{C}_n\rho)f := \frac{\int \rho(dx)\, g(x,Y_n)\, f(x)}{\int \rho(dx)\, g(x,Y_n)},$$

for any probability measure ρ on $(E,\mathcal{E})$ and any measurable bounded function f. When applied to the measure π_{n-1}, P uses the dynamics of the underlying Markov chain to "predict" X_n given the observation history $Y_1, \ldots, Y_{n-1}$, namely,

$$\mathsf{P}\pi_{n-1} = \mathbf{P}(X_n \in \cdot\,|Y_1, \ldots, Y_{n-1}).$$

Then, C_n "corrects" the predictive measure using the observation at time n, that is, it weights the measure $\mathsf{P}\pi_{n-1}$ by the likelihood function $x \to g(x, Y_n)$.

The recursive nature of the filter plays a crucial role in practice, as it allows the computations to be implemented on-line over a long time horizon. In practice, however, the optimal filter is almost never directly computable: it requires the propagation of an entire conditional distribution, which generally does not admit any efficiently computable sufficient statistics.

The practical implementation of nonlinear filtering was therefore long considered to be intractable until the discovery of a class of surprisingly efficient sequential Monte Carlo algorithms, known as *particle filters*, for approximating the filter. The simplest and most famous such algorithm is the *sequential importance resampling* (SIR) particle filter introduced by Gordon, Salmond and Smith in 1993 [28]. This algorithm simply inserts a random sampling step into the Bayes recursion and approximates the filter π_n by the resulting empirical measure $\hat{\pi}_n$. That is,

$$\hat{\pi}_{n-1} \xrightarrow{\text{prediction}} \mathsf{P}\hat{\pi}_{n-1} \xrightarrow{\text{sampling}} \mathsf{S}^N\mathsf{P}\hat{\pi}_{n-1} \xrightarrow{\text{correction}} \hat{\pi}_n := \mathsf{C}_n\mathsf{S}^N\mathsf{P}\hat{\pi}_{n-1},$$

where S^N is the sampling operator that replaces whatever measure it is applied to with its empirical measure with N independent Monte Carlo samples or *particles*, namely,

$$\mathsf{S}^N\rho := \frac{1}{N}\sum_{i=1}^{N}\delta_{X(i)}, \qquad X(1),\ldots,X(N) \text{ are i.i.d. samples with distribution } \rho,$$

where δ_x denotes the Dirac measure with mass located at x. It is not difficult to show that this gives rise to a standard Monte Carlo error (cf. Section 3.3.1)

$$\sup_{|f|\leq 1} \mathbf{E}\,|\pi_n f - \hat{\pi}_n f| \leq \frac{C}{\sqrt{N}},$$

which converges to zero in the limit for N that goes to infinity.

1.2 Filter stability

It turns out that in order to properly understand how the Monte Carlo approximation of the filter recursion behaves, we need to understand the behavior of the filter distribution itself. In fact, as shown in Section 3.3.1, a simple analysis that focuses on the filter recursion $\pi_n = \mathsf{C}_n\mathsf{P}\pi_{n-1}$ alone would yield that the constant C in the previous bound growths exponentially with time n, which is what we would expect at first as the SIR particle filter adds an approximation step (represented by the sampling operator S^N) to each iteration of Bayes formula. If the quality of the estimate given by particle filters were really to deteriorate with time, then particle filters would be totally useless in most practical applications, where one is interested in obtaining reliable estimates at *any* time. However, a deeper analysis that takes also into account also the probabilistic structure of the filter distribution π_n yields that the constant C in the previous bound does not depend on time, so that particle filters can indeed function in an on-line fashion.

Del Moral and Guionnet in 2001 [15] were the first to realize that the so-called *stability* property of nonlinear filters can be use as a dissipation mechanism for the approximation error of the SIR particle filter. Roughly speaking, filter stability says that the filter forgets its initial condition as times goes on, something like

$$\mathbf{P}(X_n \in \cdot\,|X_0 Y_1,\ldots,Y_n) \approx \mathbf{P}(X_n \in \cdot\,|Y_1,\ldots,Y_n) \qquad \text{for } n \text{ large enough.}$$

This property represents a weak form of conditional independence in time: as n increases, X_n becomes "close" to be conditionally independent of X_0 given the observation history $Y_1, \ldots, Y_n$ (different notions of "closeness" are considered in this thesis, cf. Chapter 3 and Chapter 7).

From a practical perspective, the fact that the filter is insensitive to the knowledge of the initial condition can be exploited to prove that approximation errors committed by particle filters at each time step do not accumulate over time. It turns out that the sampling step introduced at each iteration of Bayes formula is precisely the key mechanism that allows particle filters to exploit filter stability and to yield time-uniform error bounds.

As the error they commit is uniform in time, particle filters have proved to perform extraordinarily well in many classical applications such as target tracking, speech recognition, and finance [8].

1.3 Curse of dimensionality

Despite their widespread success, particle filters have nonetheless proved to be essentially useless in truly complex data assimilation problems. The reason for this, long known to practitioners, has only recently been subjected to mathematical analysis in the work of Bickel *et al.* [4, 47]. Roughly speaking, the constant C in the above bound, while independent of time n, must typically be exponential in the number of degrees of freedom of the model. This curse of dimensionality does not affect most classical tracking problems, where the dimension of the state space $E \times F$ where the model $(X_n, Y_n)_{n\geq 0}$ lives is typically of order unity. If we want to track the location of a boat, for instance, then we can take $E = \mathbb{R}^2$ (analogously, $F = \mathbb{R}^2$), which we interpret as a two dimensional space (as the motion of the boat has two degrees of freedom). On the other hand, the curse of dimensionality becomes absolutely prohibitive in large-scale data assimilation problems such as weather forecasting and oceanography, where model dimensions of order 10^7 are routinely encountered [1].

The curse of dimensionality of particle filters is a consequence of the general fact that in high dimension probability measures tend to be singular, that is, they tend to put mass on different portions of the space. The problem appears even in a single iteration of the SIR algorithm, and it is due to the correction step performed by the operator C_n: in high dimension, typical samples coming from a measure ρ have small likelihood under the measure $\mathsf{C}_n\rho$, as illustrated in Figure 1.2. Hence, in high dimension already the empirical measure $\hat{\pi}_1$ has a small fraction of particles that meaningfully approximate the filter distribution π_1 (cf. Section 3.3.3).

While this phenomenon is now fairly well understood, there exists no rigorous approach to date for alleviating this problem [3, 60]. Practical data assimilation in high-dimensional models is therefore generally performed by means of *ad-hoc* algorithms, frequently based on (questionable) Gaussian approximations, that possess limited theoretical justification [34, 37, 1].

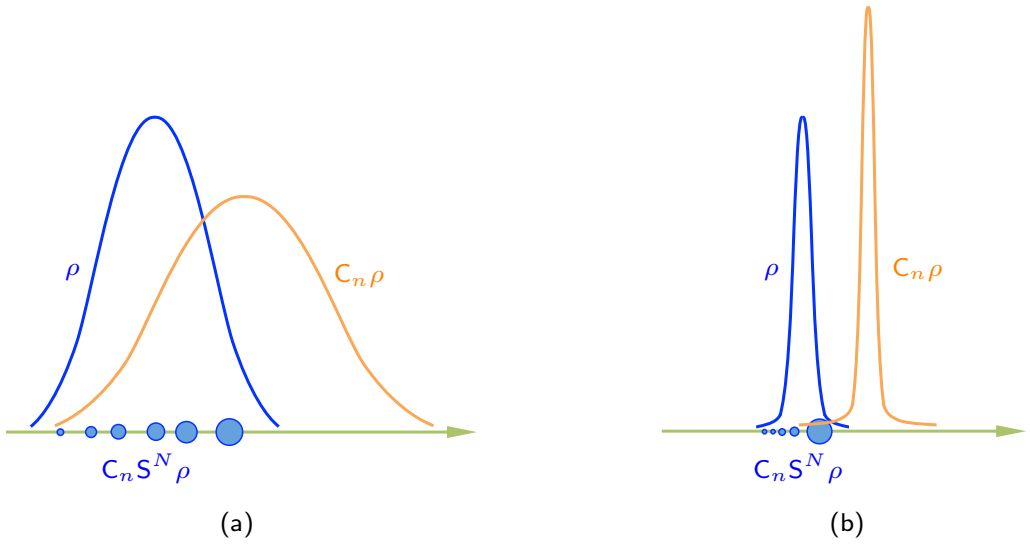

Figure 1.2: Illustration of the curse of dimensionality in a typical iteration of the SIR particle filter. (a) Probability measures in low dimension. (b) Probability measures in high dimension (low-dimensional representation). Each sample X from ρ is represented by a blue ball whose size is proportional to the likelihood $g(X, Y_n)$. In high dimension ρ and $\mathsf{C}_n\rho$ tend to put mass on different portions of the space. For this reason, in high dimension only a small fraction of the samples coming from ρ has a relevant likelihood with respect to the observation Y_n.

1.4 Fundamental obstacle in high dimension?

One of the main contribution of this thesis is to show that there is no fundamental obstacle to particle filtering in high dimension. We propose the first algorithm that can avoid the curse of dimensionality, and we develop a general framework that encompasses a novel philosophy behind filtering in high dimension. From a practical point of view, the framework that we propose provides a principled approach to design new algorithms for high-dimensional applications, where the current state of the art relies exclusively on heuristics. From a theoretical point of view, it the first time that ideas and tools from statistical mechanics are shown to play a fundamental role in the analysis of filtering models.

Before discussing the key elements that constitute our framework, we present an example that illustrates how it is possible to overcome the curse of dimensionality in a trivial setting. This example sets the direction to follow for the development of our theory.

Let $V = \{1, \dots, d\}$ be a finite index set, and for each $v \in V$ let $(X_n^v, Y_n^v)_{n\geq 0}$ be a hidden Markov model of the type being considered so far, which takes values in a measurable space $(E^v \times F^v, \mathcal{E}^v \otimes \mathcal{F}^v)$. Assume that the chains forming this collection are independent, and consider the hidden Markov model $(X_n, Y_n)_{n\geq 0}$ with $X_n = (X_n^v)_{v\in V}$ and $Y_n = (Y_n^v)_{v\in V}$. This dependency structure is illustrated in Figure 1.3.

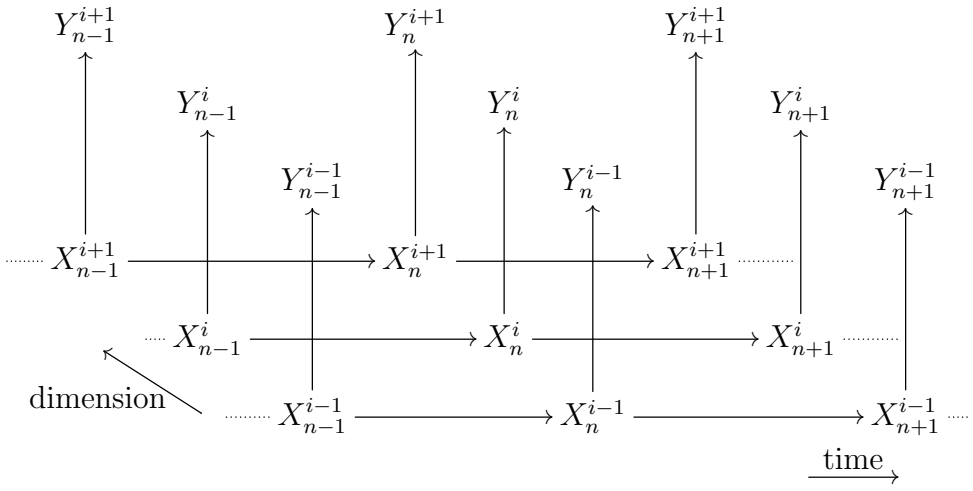

Figure 1.3: Dependency graph of a (trivial) high-dimensional filtering model.

This model clearly defines a (trivial) high-dimensional model, where the dimension is d, the number of independent chains being considered. From the theory of Bickel *et al.* [4, 47] we know that the SIR particle filter fails miserably when applied to this model, requiring a number of particles N that is exponential in d. However, in this case one can surmount this problem in a trivial fashion: as each of the coordinates of the high-dimensional model is independent, one can simply run an independent SIR filter in each coordinate. It is evident that the local error of this algorithm (that is, the error of the marginal of the filter in each coordinate) is, by construction, independent of the model dimension d. In this sense, this trivial model shows that it is indeed possible to filter very efficiently regardless of the ambient dimension (though not with the SIR particle filter, which fails spectacularly).

In the literature there is the widespread belief that filtering in high dimension is possible only if the high-dimensional model being considered lives in a low-dimensional manifold (see [10] for instance). The trivial example that we just considered, however, clearly contradicts this idea, as there is no low-dimensional structure: as the chains are independent, the global dimension is the full model dimension d. The reason why we can deal efficiently with this high-dimensional system is the fact that the model is *locally* low-dimensional (the local dimension being 1, as each coordinate is completely independent from the others), and the fact that we are interested in local errors (marginals of the filtering distribution on spatial regions of a fixed size), as opposed to the global measure of error usually considered in the literature for this type of problems.

1.5 Decay of correlations and localization

While the trivial model previously introduced does not have any practical relevance, we would like to extend the main ideas that guided our discussion in that case to

nontrivial models that are of genuine practical interest. Several fundamental questions arise immediately.

1. What sort of filtering models are natural to investigate in high dimension?
2. What sort of mechanism might allow to surmount the curse of dimensionality?
3. How can such a mechanism be exploited algorithmically?

We aim to address each of these questions in this thesis. Presently we provide an informal discussion that is instrumental to describe the main contribution of our work.

1. What filtering models are natural to investigate in high dimension? The local algorithm proposed to analyze the trivial model above (i.e., running the SIR particle filter to each chain separately) was made possible because the components of that model are truly independent. When this is not the case, we cannot run independent particle filters in each dimension as all the dimensions are coupled by the dynamics of the model. We must therefore introduce a general class of nontrivial models in which the above intuition can nonetheless be implemented.

In most data assimilation problems, the high-dimensional nature of the model is essentially due to its spatial structure: the aim of the problem is to track the dynamics of a random field (for example, the atmospheric pressure and temperature fields in the case of weather forecasting). We therefore take as a starting point the notion that the coordinates X_n^v, Y_n^v ($v \in V$) of our hidden Markov model are indexed by a large graph $G = (V, E)$ that represents the spatial degrees of freedom of the model. That is, we consider the case $(E, \mathcal{E}) = (\bigtimes_{v\in V} E^v, \bigotimes_{v\in V} \mathcal{E}^v)$ and $(F, \mathcal{F}) = (\bigtimes_{v\in V} F^v, \bigotimes_{v\in V} \mathcal{F}^v)$. It is of course not reasonable to expect that the dynamics at each spatial location is independent, as was assumed in the trivial model previously discussed. On the other hand, dynamics of spatial systems is typically local in nature: the dynamics at a spatial location depends only on the states at locations in a neighborhood. Moreover, the observations are typically local in the sense that (a subset of) spatial locations are observed independently. The dependency structure of this type of models is illustrated in Figure 1.4.

These local filtering models are prototypical of a broad range of high-dimensional filtering problems, and they provide the basic framework for our main result. They arise naturally in numerous complex and large-scale applications, including percolation models of disease spread or forest fires, freeway traffic flow models, probabilistic models on networks and large-scale queueing systems, and various biological, ecological and neural models. Moreover, local Markov processes of this type arise naturally from finite-difference approximation of stochastic partial differential equations, and are therefore in principle applicable to a diverse set of data assimilation problems that arise in areas such as weather forecasting, oceanography, and geophysics (cf. Section 4.4.4).

2. What mechanism can allow to surmount the curse of dimensionality? While the law of the model at each spatial location is no longer independent as in the

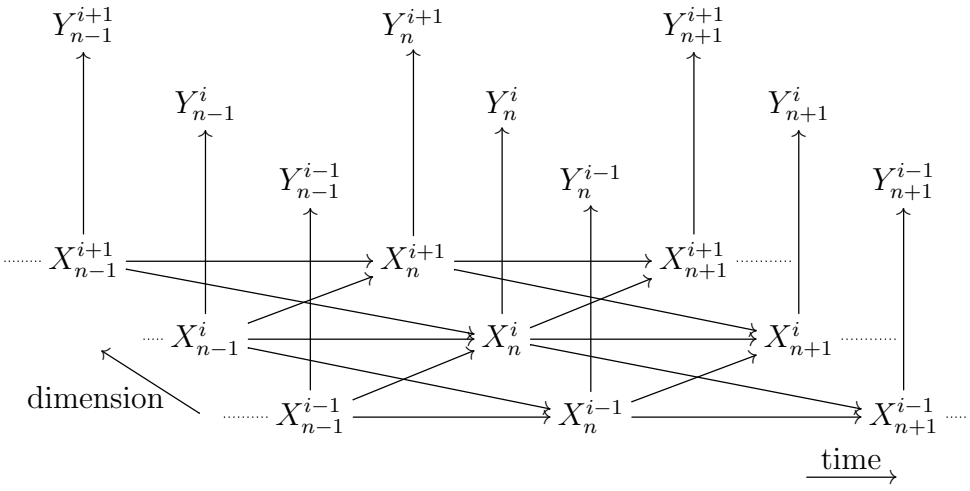

Figure 1.4: Dependency graph of a high-dimensional filtering model of the type considered in this thesis.

trivial model of the previous section, large-scale interacting systems can nonetheless exhibit an approximate version of this property: this is the *decay of correlations* phenomenon that has been particularly well studied in statistical mechanics (see, e.g., [27]). Informally speaking, while the states (X^v_n, Y^v_n) and (X^w_n, Y^w_n) at two sites $v, w \in V$ are probably quite strongly correlated when v and w are close together, one might expect that (X^v_n, Y^v_n) and (X^w_n, Y^w_n) are nearly independent when v and w are far apart as measured with respect to the natural distance d in the graph G (that is, $d(v, w)$ is the length of the shortest path in G between $v, w \in V$). The idea is that due to the decay of correlations, also this type of model is locally low-dimensional, in the sense that the conditional distribution of each coordinate only needs to be updated by observations in a neighborhood whose size is independent of the ambient dimension. That is,

$$\mathbf{P}(X^v_n \in \cdot\,|Y_1, \ldots, Y_n) \approx \mathbf{P}(X^v_n \in \cdot\,|Y^w_1, \ldots, Y^w_n, d(v, w) \leq b),$$

for b large enough. Roughly speaking, the "local dimension" of the model is the number of coordinates in a ball whose radius is the *correlation length* of the filtering distribution.

3. How can such a mechanism be exploited algorithmically?
Both filter stability and decay of correlations are probabilistic properties of the filter distribution itself: filter stability represents a weak form of conditional independence in time, and the decay of correlations property represents a weak form of conditional independence in space (model dimension). As already mentioned, the sampling step added to the original filter recursion is the key to exploit algorithmically filter stability and get particle filters that yield time-uniform error bounds. One of the main goal of this thesis is to show that proper forms of *localization* of the filter recursion can be used to exploit algorithmically the decay of correlations property and to design *local* particle filters that yield error bounds that are uniform both in time and in space.

As mentioned above, the curse of dimensionality of particle filters is essentially due to the fact that probability measures tend to be singular in high dimension. However, while this is definitely what happens if we consider a high-dimensional model as a whole, if the decay of correlations property holds then it should be possible to localize the model and work with local low-dimensional portions of it. As the problem comes from the correction step of the filter recursion, what really matters is the dimension of the observations (cf. Section 3.3.2), and it makes sense to introduce a localization step immediately before the operator C_n so that the model can behave as "local" for the sake of likelihood-reweighting.

A speculative back-of-the-envelope computation explains how this might work. Due to the decay of correlations, the conditional distribution of the site X_n^v given the new observation Y_n should not depend significantly on observations Y_n^w at sites w distant from v. Suppose we can develop a local particle filtering algorithm that at each site v only uses observations in a local neighborhood K of v to update the filtering distribution. As we have now restricted to observations in K, the sampling error (the *variance*) at each site will be exponential only in $\operatorname{card} K$ rather than in the full dimension $\operatorname{card} V$. On the other hand, the truncation to observations in K is only approximate: the decay of correlations property suggests that the *bias* introduced by this truncation should decay exponentially in $\operatorname{diam} K$. Therefore,

$$\text{error} = \text{bias} + \text{variance} \approx e^{-\operatorname{diam} K} + \frac{e^{\operatorname{card} K}}{\sqrt{N}}.$$

If the size of the neighborhoods K is chosen so as to optimize the error, then the resulting algorithm is evidently consistent (with a slower convergence rate than the standard $1/\sqrt{N}$ Monte Carlo rate: this is likely unavoidable in high dimension) with an error bound that is independent of the model dimension $\operatorname{card} V$.

So, the general idea of local particle filters is that one should introduce a spatial regularization step into the filtering recursion that enables local sampling. While these regularizations introduce some bias to ordinary particle filters, they largely reduce their variance, and it is exactly the bias-variance tradeoff that emerges that can be used to overcome the curse of dimensionality.

In this thesis we develop two localization procedures that aim at implementing this idea.

1. Using independence: block particle filter

The most natural way to localize the filter recursion is to marginalize it. In this thesis we analyze the *block particle filter* that we define iteratively as:

$$\begin{aligned}
\hat{\pi}_{n-1} &\xrightarrow{\text{prediction}} \mathsf{P}\hat{\pi}_{n-1} &&\xrightarrow{\text{sampling}} \mathsf{S}^N\mathsf{P}\hat{\pi}_{n-1} \\
&\xrightarrow{\text{blocking}} \mathsf{B}\mathsf{S}^N\mathsf{P}\hat{\pi}_{n-1} &&\xrightarrow{\text{correction}} \hat{\pi}_n := \mathsf{C}_n\mathsf{B}\mathsf{S}^N\mathsf{P}\hat{\pi}_{n-1},
\end{aligned}$$

where B is an operator that projects a measure to the product of its marginals over a certain partition $\mathcal{K}$ of "blocks" of the index set V, that is,

$$\mathsf{B}\rho := \bigotimes_{K\in\mathcal{K}} \mathsf{B}^K\rho,$$

where for any measure ρ on $(E, \mathcal{E})$ and $J \subseteq V$ we denote by $\mathsf{B}^J\rho$ the marginal of ρ on $(\bigtimes_{v\in J} E^v, \bigotimes_{v\in J} \mathcal{E}^v)$.

This algorithm captures the main intuition that motivated our discussion on local algorithms: choosing $\mathcal{K} = \bigcup_{v\in V}\{v\}$, in fact, the block particle filter reduces to applying the SIR particle filter independently to each of the components constituting the model (which, of course, introduces a bias unless the components are independent).

In Chapter 4 we show that this local particle filter surmounts the key obstacle in high dimension by providing local estimates that are uniform in time and that do not depend on the ambient dimension.

2. Using conditional independence: localized Gibbs sampler particle filter The block particle filter possesses some inherent limitations as it can only provide spatially inhomogeneous approximations of the filter distribution. In fact, at each iteration the algorithm projects the approximated filter measure into the product of its marginals over a given (fixed) partition of the environment space.

In order to address this deficiency at a fundamental level, we consider a regularization that aims at projecting probability measures to the class of Markov random fields (of a certain interaction neighborhood), instead of projecting them to the class of distributions that are independent across subsets of coordinates, as in the block particle filter. This is precisely the idea that animates the *localized Gibbs sampler particle filter*. Heuristically, this algorithm can be described as follows

$$\begin{array}{ccccc} \hat{\pi}_{n-1} & \xrightarrow{\text{prediction}} & \mathsf{P}\hat{\pi}_{n-1} & \xrightarrow{\text{projection}} & \mathsf{MP}\hat{\pi}_{n-1} \\ & \xrightarrow{\text{correction}} & \mathsf{C}_n\mathsf{MP}\hat{\pi}_{n-1} & \xrightarrow{\text{sampling}} & \hat{\pi}_n := \mathsf{S}^N\mathsf{C}_n\mathsf{MP}\hat{\pi}_{n-1}, \end{array}$$

where M is an operator that projects a measure to the class of Markov random fields of order b, that is,

$$(\mathsf{M}\rho)(X^v \in A | X^{V\setminus\{v\}} = x^{V\setminus\{v\}}) = \rho(X^v \in A | X^{N_b(v)\setminus\{v\}} = x^{N_b(v)\setminus\{v\}})$$

for every $v \in V$, where $N_b(v) := \{v' \in V : d(v, v') \leq b\}$.

As shown in Chapter 5, by sampling locally in each dimension rather than globally over all dimensions, the localized Gibbs sampler particle filter implements a sort of "resampling in space." In this sense, the mechanism through which this algorithm exploits the decay of correlations property to provide spatially-uniform error bounds resembles the analogous mechanism that allows the SIR particle filter to exploit filter stability and provide time-uniform error bounds.

While a complete analysis of this algorithm is still missing, we prove a one-step error bound for the bias term that illustrates the way this algorithm can provide spatially-uniform error bounds.

1.6 Comparison theorems for Gibbs measures

At this point it is not at all clear what sort of mathematical tools are needed to make the speculative ideas discussed so far precise. In fact, the rigorous implementation

of these ideas requires the introduction of a mathematical machinery that has not previously been applied in the study of nonlinear filtering.

As the trivial example previously introduced illustrates, in order to describe effectively filtering problems in high dimension it is necessary to perform a local analysis: we want to look at a local measure of the error, and we want to be able to perform the analysis using local quantities of the model. As any approximation of practical utility in high dimension must yield error bounds that do not grow, or at least grow sufficiently slowly, in the model dimension $\operatorname{card} V$, we seek for quantitative methods that allow to establish dimension-free bounds on high-dimensional probability distributions.

A general method to address precisely this problem is the Dobrushin comparison theorem that was developed by Dobrushin in the context of statistical mechanics [18, Theorem 3]. In the approach pioneered by Dobrushin, Lanford, and Ruelle, a high-dimensional (possibly infinite) system of interacting random variables is defined by its local description: for finite sets of sites $J \subset V$, the conditional distribution $\rho(X^J \in \cdot\,|X^{V\setminus J} = x^{V\setminus J})$ of the configuration in J is specified given that the variables outside J are frozen in a fixed configuration $x^{V\setminus J}$ (we write $x^K = (x^k)_{k\in K}$ for $K \subseteq V$). The model ρ is then defined as a probability measure (called a *Gibbs measure*) that is compatible with the given system of local conditional distributions.

The Dobrushin comparison theorem is a tool to bound the total variation difference between marginals of Gibbs measures in terms of their local conditional distributions. This tool is what allow us to characterize the crucial way in which the decay of correlations property enters the local analysis of particle filters in our framework.

Despite being a powerful tool, the Dobrushin comparison theorem requires the validity of restrictive assumptions, and for most models this fact poses a major limitation on the applicability of the theorem.

One of the contribution of this thesis is to develop a more general and flexible machinery that allows us to get more powerful results. By relying on the Dobrushin-Shlosman [17] and Weitz [64] conditions for uniqueness of Gibbs measures, instead of the Dobrushin condition employed by the original comparison theorem, the new comparison theorems that we develop in Chapter 6 provide more flexible tools to analyze the behavior of algorithms in larger regions of the natural parameter space, and are of independent interest in statistical mechanics for the analysis of Gibbs measures.

The novel toolbox is used to extend qualitatively the analysis of the block particle filter that we initially obtain using the original Dobrushin comparison theorem.

In order to prove the new comparison theorems we develop a methodology that exploits the connection with a certain type of Markov chains called Gibbs samplers. The general framework behind our proofs represents a novel contribution in the connection between static Gibbs measures and dynamical Gibbs samplers.

1.7 Filtering in infinite dimension

Alongside with the quantitative investigation of Gibbs measures and its connection to filtering algorithms in high dimension, this thesis also deals with the qualitative understating of conditional ergodicity of Gibbs measures and its connection to the theory of filtering in infinite dimension.

As previously discussed, both filter stability and decay of correlations are probabilistic properties of the filter distribution that play a crucial role in the development of particle filters. At first sight, both properties might also seem natural. If we take filter stability, for instance, it is often the case that the underlying chain $(X_n)_{n\geq 0}$ itself is stable, namely,

$$\mathbf{P}(X_n \in \cdot\,|X_0) \approx \mathbf{P}(X_n \in \cdot\,) \qquad \text{for } n \text{ large enough,}$$

and it seems highly likely that if this is the case then also the filter should forget its initial condition. Moreover, it seems natural that as time n increases the initial knowledge of X_0 is superseded by the information contained in the observations $Y_1, \ldots, Y_n$, so that eventually X_0 does not affect the filter. However, neither of these two intuitions is always true.

Understanding the general assumptions that guarantee the inheritance of stability from the underlying chain to the filter distribution has been a longstanding problem dating back to the work of Blackwell in 1957 [5] and Kunita in 1971 [33], and it is related to many areas of probability theory, far beyond the algorithmic setting considered in this thesis [63, 59].

A general qualitative theory that exhaustively characterizes this phenomenon has recently been developed by van Handel in 2009 [57], where it is shown that if the Markov chain $(X_n, Y_n)_{n\geq 0}$ is stable (in a certain total variation sense), and if a mild non-degeneracy condition holds for density of the law of Y_k given X_k for each $k \geq 0$ (essentially requiring the presence of some noise in the observations), then the filter $\mathbf{P}(X_n \in \cdot\,|Y_1, \ldots, Y_n)$ is also stable. This result represents a milestone in the theory of nonlinear filtering, settling a long dispute in the field.

However, while this result holds in a very general setting and there is no explicit mention of dimensionality, in practice it can only be applied to finite-dimensional systems. In fact, on the one hand, if the underlying signal $(X_n)_{n\geq 0}$ has an infinite-dimensional state space, then the ergodicity assumption in total variation can not be satisfied; on the other hand, if the observations $(Y_n)_{n\geq 0}$ are infinite-dimensional, then the non-degeneracy condition can not hold. In [52] it has been shown that the infinite dimensionality of the underlying signal is not a fundamental issue, and that the main filter ergodicity result in [57] still holds true, either upon working with a local notion of convergence in total variation, or upon doing the analysis in weak convergence, which embodies a form of locality in itself. However, this later development still requires the same global non-degeneracy assumption as in [57], which essentially restricts the scope of the theory to finite-dimensional observations.

One of the contribution of this thesis is to develop the first results in filtering with infinitely many observations, and to show that in this setting completely new

phenomena can appear. For instance, in Chapter 7 we show that we can have a completely ergodic infinite-dimensional model (X, Y), where the underlying system X is a collection of independent random variables and the structure of the observation Y is local, and still it is possible for the conditional distribution $\mathbf{P}(X \in \cdot|Y)$ to display a phase transition in the signal-to-noise ratio (see Theorem 7.7 and Example 7.17). That is, as we condition on the observations there is a threshold showing up such that if the signal-to-noise ratio is below it, then the conditional distribution is unique; else, the conditional distribution is not unique. This example shows that while the ergodicity of the underlying process can be localized so to recover the powerful general result as in [57], localizing the non-degeneracy in the conditional law of the observations does not help.

Far from being a theoretical point, the understanding of filtering theory in infinite dimension is crucial for the development of particle filters that can work in practical applications. In fact, it is well know that the *qualitative* understanding of infinite-dimensional models is directly related to the *quantitative* understanding of finite-dimensional models (see [50] and [39] for instance).

1.8 Outline of the thesis

This thesis consists of 7 chapters and 4 appendices.

Chapter 1 is the introduction.

Chapter 2 contains a collection of results that are used repeatedly throughout this thesis. As a large portion of this thesis deals with controlling the distance between conditional distributions in high dimension, we present a few elementary lemmas that serve this purpose, along with the main tool that is used in our proofs—the Dobrushin comparison theorem from statistical mechanics. We also give a brief overview of Monte Carlo methods, as they are needed to describe the algorithms presented in the next chapters. The goal of this chapter is to provide the necessary tools that are needed in the remaining of this thesis, along with establishing the notation that is used throughout.

Chapter 3 provides an introduction to the classical theory of nonlinear filtering and sequential Monte Carlo algorithms known as particle filters. Particle filters are discussed in the light of the curse of dimensionality phenomenon. First, the *sequential importance sampling* (SIS) algorithm is introduced, and it is showed how it suffers from the curse of dimensionality with respect to time. This issue motivates the introduction of the *sequential importance resampling* (SIR) algorithm, for which time-uniform error bounds can be proved. The notion of filter stability plays a central role in establishing bounds that do not depend on time. Nonetheless, it is showed that both algorithms still suffer from the curse of dimensionality with respect to the spatial dimension of the model. This discussion paves the way for the introduction of *local* particle filters that is developed in the next two chapters. The treatment of the material presented in this chapter is inspired by [55] and [8].

Chapter 4 introduces the *block particle filter* and shows how this algorithm overcomes the curse of dimensionality by yielding errors bounds that are uniform both

in time and in space. More generally, this chapter introduces the class of high-dimensional filtering models that we consider in this thesis, and it illustrates how the Dobrushin comparison theorem can be used to perform a local analysis in these models. Emphasis is given to the decay of correlations property, which is seen to be the key to establish spatially-uniform error bounds, thus representing the spatial counterpart of filter stability. It is showed how decay of correlations can be exploited algorithmically by introducing a regularization step (marginalization over non-overlapping blocks) to the basic formulation of the SIR algorithm. The analysis of the block particle filter is instrumental to developing a general framework that can encompass other algorithms (that is, other forms of regularization), such as the one proposed in the next chapter. To facilitate the reading, the proofs of the results presented in this chapter are included in Appendix A. This chapter is based on the paper [40].

Chapter 5 introduces the *localized Gibbs sampler particle filter*, another local algorithm that aims at exploiting the decay of correlations property of filtering models through a form of regularization based on the notion of conditional independence (rather than on the notion of independence, as for the block particle filter). While a complete analysis of this algorithm is still missing, we prove a one-step error bound that illustrates how this algorithm provides spatially homogenous approximations of the filter distribution, hence overcoming the main drawback of the block particle filter (the proof is included in Appendix B). The analysis of this algorithm prompts for the investigation of the decay of correlations in general Markov Chain Monte Carlo methods, and new challenges arise in this context. The material presented in this chapter is new and has not been submitted to publication yet.

Chapter 6 is devoted to establishing new comparison theorems for Gibbs measures that extend the applicability of the original Dobrushin comparison theorem to larger regions of the phase space. The proof of these results (contained in Appendix C) is part of a more general framework that is developed to analyze the convergence behavior of Gibbs samplers, a particular class of Markov chains. As an application, the new comparison theorems are used to improve *qualitatively* the analysis of the block particle filter given in Chapter 4 to handle scenarios where ergodicity in space and in time are treated on a different footing. This chapter is based on the paper [41].

Chapter 7 presents some of the first results in the theory of filtering with infinitely-many observations. The focus of this chapter is complementary to the *quantitative* framework previously analyzed in this thesis, mostly in the realm of algorithms. Now we are interested in the fundamentals of filtering theory in infinite dimension, and filter stability and decay of correlations are analyzed *qualitatively* in models with infinitely-many degrees of freedom. We show that completely new phenomena appear in this setting: contrarily to the finite-dimensional case, inheritance of ergodicity can undergo a phase transition in the signal-to-noise ratio. We refer to Appendix D for the proofs of the results presented in this chapter. The material of this chapter is taken from the paper [42], which further develops this set of ideas by yielding, for instance, conditions that guarantee the inheritance of ergodicity.

Chapter 2

Preliminaries

This chapter is devoted to introducing some elementary concepts and facts in probability theory that will be needed in what follows. As this thesis ultimately deals with conditioning, large focus is given to the notion of conditional expectations and conditional distributions, along with some of their basic properties. Since we will be mostly concerned with high-dimensional distributions, we present a collection of tools to control their distances. Emphasis is also given to the Monte Carlo paradigm, which is the backbone of the first part of this thesis. The material is presented in a streamlined manner, and no attempt is made at developing a systematic treatment. This chapter also serves to set the notation being adopted in this thesis.

We assume that the reader is already familiar with measure-theoretic probability theory at the level of an introductory class on the subject. We refer to [65] for an agile and beautiful introduction to this material, and to [11] for a comprehensive and systematic treatment of it.

2.1 Notation and conventions

We begin by establishing some notations and conventions that will be used throughout.

A function from a measurable space $(E, \mathcal{E})$ to $\bar{\mathbb{R}} := [-\infty, +\infty]$, or a subset of it, is $\mathcal{E}$-*measurable* it if is measurable relative to $\mathcal{E}$ and the Borel σ-algebra on $\bar{\mathbb{R}}$. We write $f \in \mathcal{E}$ to mean that the function f is $\mathcal{E}$-measurable. We write $\mathbf{1}_A$ for the indicator function on the event $A \in \mathcal{E}$. We say that a function is *positive* if it takes values in $\bar{\mathbb{R}}_+ := [0, +\infty]$.

When we say that X is a $(E, \mathcal{E})$-valued random variable with distribution μ, we mean that there is a probability space $(\Omega, \mathcal{H}, \mathbf{P})$ in the background so that X is a random variable taking values on the measurable space $(E, \mathcal{E})$ and

$$\int \mathbf{P}(d\omega)\, f(X(\omega)) \equiv \mathbf{E}\, f(X) = \mu f \equiv \int \mu(dx)\, f(x)$$

for any positive $\mathcal{E}$-measurable function f. If X has distribution μ, we write $X \sim \mu$. Given a random variable X, we denote by σX the σ-algebra generated by it.

To keep the use of parentheses at minimum, we write $\mathbf{E}\,XY$ to mean $\mathbf{E}(XY)$. To avoid pedantry, whenever easily inferred by the context, we will often not specify the domain where functions are defined, the σ-algebras where events live, and the σ-algebras involved with the definition of measurable functions. For instance, we will often say that a probability measure μ on a measurable space $(E, \mathcal{E})$ is defined by $\mu(dx)$, $\mu(A)$, or μf, without mentioning that this definition has to hold, respectively, for each $x \in E$, each $A \in \mathcal{E}$, or for each positive $\mathcal{E}$-measurable function f. Often we will also say that μ is a probability measure on E, when we really mean $(E, \mathcal{E})$.

2.2 Conditioning and Bayes formula

In this thesis we will be primarily interested in the behavior of conditional expectations and conditional distributions. We presently recall some of the main definitions and properties that will be needed in what follows. As a large part of this thesis is devoted to Monte Carlo approximations, emphasis is given to the role of random variables. As such, results in this section are mainly phrased in terms of random variables, distribution of random variables, and σ-algebras generated by random variables, rather than in terms of probability measures and generic σ-algebras.

Definition 2.1 (Conditional expectation)**.** *Let X be a $\bar{\mathbb{R}}$-valued random variable, and let Y be a $(F, \mathcal{F})$-valued random variable. The* conditional expectation *of X given Y is any random variable of the form $h(Y)$, where h is a $\bar{\mathbb{R}}$-valued $\mathcal{F}$-measurable function, such that the following holds for any positive $\mathcal{F}$-measurable function f:*

$$\mathbf{E}\, h(Y) f(Y) = \mathbf{E}\, X f(Y).$$

We use the notation $\mathbf{E}(X|Y)$ to indicate any such random variable $h(Y)$. We also write $\mathbf{E}(X|Y = y)$ to mean the value that any such function h takes at $y \in F$, that is, $\mathbf{E}(X|Y = y) = h(y)$ (recall that conditional expectations are defined up to almost surely equivalences).

The conditional expectation of X given Y is the function of Y that provides best estimates of X in the least square sense, as the following lemma shows.

Lemma 2.2 (Optimality of conditional expectation)**.** *Let X be a $\bar{\mathbb{R}}$-valued random variable, and let Y be a $(F, \mathcal{F})$-valued random variable. Assume that $\mathbf{E}\, X^2 < \infty$. Then, the function $y \in F \to h(y) := \mathbf{E}(X|Y = y)$ satisfies*

$$h = \arg\min_g \mathbf{E}\,(X - g(Y))^2,$$

where the minimization is with respect to all $\bar{\mathbb{R}}$-valued measurable functions g such that $\mathbf{E}\, g(Y)^2 < \infty$.

Proof. By the properties of conditional expectations we have

$$\begin{aligned}\mathbf{E}\,(X - h(Y))^2 &= \mathbf{E}\, X^2 + \mathbf{E}\, \mathbf{E}(X|Y)^2 - 2\mathbf{E}\, X \mathbf{E}(X|Y) \\ &= \mathbf{E}\, X^2 - \mathbf{E}\, \mathbf{E}(X|Y)^2 \leq \mathbf{E}\, X^2 < \infty.\end{aligned}$$

It remains to prove that for any measurable function g we have

$$\mathbf{E}\,(X - h(Y))^2 \leq \mathbf{E}\,(X - g(Y))^2.$$

For simplicity, define $H := h(Y) = \mathbf{E}[X|Y]$ and $G := g(Y)$. Then,

$$\begin{aligned}\mathbf{E}\,(X - H)^2 &= \mathbf{E}\,(X - G + G - H)^2 \\ &= \mathbf{E}\,(X - G)^2 + \mathbf{E}\,(G - H)^2 + 2\mathbf{E}((X - G)(G - H)) \\ &= \mathbf{E}\,(X - G)^2 - \mathbf{E}\,(G - H)^2 \\ &\leq \mathbf{E}\,(X - G)^2,\end{aligned}$$

where we used that, by the properties of conditional expectations,

$$\begin{aligned}\mathbf{E}\,(X - G)(G - H) &= \mathbf{E}\,\mathbf{E}((X - G)(G - H)|Y) = \mathbf{E}\,\mathbf{E}(X - g(Y)|Y)(G - H) \\ &= \mathbf{E}\,(H - G)(G - H) = -\mathbf{E}\,(G - H)^2.\end{aligned}$$

□

Let us recall the definition of transition kernels, which is instrumental for the definition of conditional distributions given immediately below.

Definition 2.3 (Transition kernel)**.** *Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be measurable spaces. Let K be a mapping from $E \times \mathcal{F}$ into $\bar{\mathbb{R}}_+$. Then, K is called a* transition kernel *from $(E, \mathcal{E})$ to $(F, \mathcal{F})$ if the following two conditions are satisfied:*

(a) the mapping $x \to K(x, B)$ is $\mathcal{E}$-measurable for every set $B \in \mathcal{F}$;

(b) the mapping $B \to K(x, B)$ is a measure on $(F, \mathcal{F})$ for every $x \in E$.

If K is a transition kernel from $(E, \mathcal{E})$ to $(F, \mathcal{F})$, μ is a probability measure on $(E, \mathcal{E})$ and f is a $(F, \mathcal{F})$-measurable function, we use the notation

$$Kf(x) \equiv K_x f \equiv \int K(x, dy)\, f(y),$$

$$\mu K f \equiv \int \mu(dx)\, K(x, dy)\, f(y).$$

Definition 2.4 (Conditional distribution)**.** *Let X be an $(E, \mathcal{E})$-valued random variable, and let Y be a $(F, \mathcal{F})$-valued random variable. A probability kernel $P : F \times \mathcal{E} \to [0, 1]$ which satisfies*

$$Pf(Y) = \mathbf{E}(f(X)|Y)$$

for every positive $\mathcal{E}$-measurable function is called the conditional distribution *of X given Y. Sometimes we also write $\mathbf{P}(X \in dx|Y)$ to mean $P(Y, dx)$.*

Remark 2.5 (Random measure)**.** *Given a measurable space $(E, \mathcal{E})$, a* random measure *μ on $(E, \mathcal{E})$ is a transition kernel from the underlying probability space $(\Omega, \mathcal{H}, \mathbf{P})$ to $(E, \mathcal{E})$. We say that a collection of random variables $X_1, \ldots, X_N$ on $(E, \mathcal{E})$ is i.i.d. coming from the random measure μ on $(E, \mathcal{E})$ if there exists a random variable Y taking values in some measurable space $(F, \mathcal{F})$ such that the following holds true for all positive $\mathcal{E}$-measurable functions $f_1, \ldots, f_N$:*

$$\mathbf{E}(f_1(X_1) \cdots f_N(X_N)|Y) = \mathbf{E}(f_1(X_1)|Y) \cdots \mathbf{E}(f_N(X_N)|Y) = \mu f_1(Y) \cdots \mu f_N(Y).$$

Given two probability measures μ and ν on a measurable space $(E, \mathcal{E})$, recall the following definitions. If for each $A \in \mathcal{E}$ such that $\nu(A) = 0$ we have $\mu(A) = 0$, then μ is said to be *absolutely continuous* with respect to ν, and we write $\mu \ll \nu$. If $\mu \ll \nu$ and $\nu \ll \mu$, then μ and ν are said to be *equivalent*, and we write $\mu \sim \nu$. If there exists $A \in \mathcal{E}$ such that $\mu(A) = 0$ and $\nu(A) = 1$, then μ and ν are said to be *mutually singular*, and we write $\mu \perp \nu$.

The following is a key result that relates probability measures that are absolutely continuous.

Theorem 2.6 (Radon-Nikodym derivative)**.** *Let X and Z be $(E, \mathcal{E})$-valued random variables with distribution μ and ν respectively. Assume that $\mu \ll \nu$. Then, there exists a positive $\mathcal{E}$-measurable function $\frac{d\mu}{d\nu}$ called the* Radon-Nikodym derivative *such that*

$$\mathbf{E}\, f(X) = \mathbf{E}\, \tfrac{d\mu}{d\nu}(Z)\, f(Z)$$

for each positive $\mathcal{E}$-measurable function f.

Proof. We refer to [65] for a proof of such result. □

The following is a key result that relates the way absolutely continuous probability measures behave under conditioning. It is one of the many forms of *Bayes formula*.

Theorem 2.7 (Bayes formula)**.** *Let X and Z be two $(E, \mathcal{E})$-valued random variables, and let Y be a $(F, \mathcal{F})$-valued random variable. Let μ be the distribution of (X, Y), and let ν be the distribution of (Z, Y), with $\mu \ll \nu$. Then, for each positive $\mathcal{E}$-measurable function f we have*

$$\mathbf{E}(f(X)|Y) = \frac{\mathbf{E}(\frac{d\mu}{d\nu}(Z, Y)\, f(Z)|Y)}{\mathbf{E}(\frac{d\mu}{d\nu}(Z, Y)|Y)}.$$

Let Q be the conditional distribution of Z given Y. Then, the conditional distribution of X given Y is given by the probability kernel P defined as

$$P(y, dz) = \frac{Q(y, dz)\, \frac{d\mu}{d\nu}(z, y)}{\int Q(y, dz)\, \frac{d\mu}{d\nu}(z, y)}.$$

Proof. We only prove the statement for conditional expectations, as the statement for conditional probabilities follows immediately by applying Definition 2.4. Fix a positive $\mathcal{E}$-measurable function f. First, we prove that

$$\mathbf{E}(\tfrac{d\mu}{d\nu}(Z, Y)\, f(Z)|Y) = \mathbf{E}(f(X)|Y)\, \mathbf{E}(\tfrac{d\mu}{d\nu}(Z, Y)|Y).$$

As the right-hand side is clearly a function of Y, by definition of conditional expectations we only need to prove that

$$\mathbf{E}\, \mathbf{E}(f(X)|Y)\, \mathbf{E}(\tfrac{d\mu}{d\nu}(Z, Y)|Y)\, g(Y) = \mathbf{E}\, \tfrac{d\mu}{d\nu}(Z, Y)\, f(Z)\, g(Y)$$

for every positive $(F, \mathcal{F})$-measurable function g. In fact, using the properties of conditional expectations and using the Radon-Nikodym theorem (Theorem 2.6) for the

random variables (X, Y) and (Z, Y) we have

$$\begin{aligned}\mathbf{E}\,\mathbf{E}(f(X)|Y)\,\mathbf{E}(\tfrac{d\mu}{d\nu}(Z,Y)|Y)\,g(Y) &= \mathbf{E}\,\mathbf{E}(f(X)|Y)\,\tfrac{d\mu}{d\nu}(Z,Y)\,g(Y)\\ &= \mathbf{E}\,\mathbf{E}(f(X)|Y)\,g(Y)\\ &= \mathbf{E}\,f(X)\,g(Y)\\ &= \mathbf{E}\,\tfrac{d\mu}{d\nu}(Z,Y)\,f(Z)\,g(Y).\end{aligned}$$

To conclude the proof we only need to prove that

$$\mathbf{E}(\tfrac{d\mu}{d\nu}(Z,Y)|Y) > 0 \qquad \mathbf{P}\text{-a.s.}$$

Using agan the Radon-Nikodym theorem we get

$$\begin{aligned}\mathbf{E}\,\mathbf{1}_{\mathbf{E}(\frac{d\mu}{d\nu}(Z,Y)|Y)=0}(Y) &= \mathbf{E}\,\tfrac{d\mu}{d\nu}(Z,Y)\,\mathbf{1}_{\mathbf{E}(\frac{d\mu}{d\nu}(Z,Y)|Y)=0}(Y)\\ &= \mathbf{E}\,\mathbf{E}(\tfrac{d\mu}{d\nu}(Z,Y)|Y)\,\mathbf{1}_{\mathbf{E}(\frac{d\mu}{d\nu}(Z,Y)|Y)=0}(Y) = 0.\end{aligned}$$

□

As an immediate application of Bayes formula we have the following lemma on the computation of conditional distributions. While this lemma could be proved using directly the definition of conditional expectation, we prove it using Bayes formula, as it is representative of the way Bayes formula will often be used in this thesis.

Lemma 2.8 (Computation of conditional distributions)**.** *Let X be an $(E, \mathcal{E})$-valued random variable, and let Y be a $(F, \mathcal{F})$-valued random variable such that for each positive $(E \times F, \mathcal{E} \otimes \mathcal{F})$-measurable function f we have*

$$\mathbf{E}\,f(X,Y) = \int \rho(dx)\,\lambda(dy)\,\gamma(x,y)\,f(x,y),$$

where ρ and λ are probability measures on $(E, \mathcal{E})$ and $(F, \mathcal{F})$ respectively, and γ is a strictly positive $\mathcal{E} \otimes \mathcal{F}$-measurable function. Then, the conditional distribution of X given Y is given by the probability kernel P defined as

$$P(y, dx) = \frac{\rho(dx)\,\gamma(x,y)}{\int \rho(dx)\,\gamma(x,y)}.$$

Proof. Define the following two probability measures on $(E \times F, \mathcal{E} \otimes \mathcal{F})$:

$$\begin{aligned}\mu(dx, dy) &:= \rho(dx)\,\lambda(dy)\,\gamma(x,y),\\ \nu(dx, dy) &:= \rho(dx)\,\lambda(dy).\end{aligned}$$

Clearly $\mu \ll \nu$ and $\frac{d\mu}{d\nu} = \gamma$. By definition μ is the distribution of (X, Y). Let Z be an $(E, \mathcal{E})$-valued random variable such that ν is the distribution of (Z, Y). By independence we immediately find that the conditional distribution of Z given Y is given by the probability kernel Q defined as

$$Q(y, dz) = \rho(dz).$$

Then, by Bayes formula (Theorem 2.7) the result follows immediately. □

2.3 Distances between probability measures

In this thesis we will face the problem of measuring and controlling the distance between probability measures. To this end, we currently introduce the two main notions of distance that we will consider, along with some elementary lemmas on their behavior.

Let $(E, \mathcal{E})$ be a measurable space, and let μ and ν be two (possibly random) probability measures on it. We define the *total variation* distance between μ and ν as

$$\|\mu - \nu\| := \sup_{f \in \mathcal{E}: \|f\|_\infty \leq 1} |\mu f - \nu f|,$$

where $\|f\|_\infty := \sup_{x \in E} |f(x)|$. We will also need the following distance between probability measures:

$$|||\mu - \nu||| := \sup_{f \in \mathcal{E}: \|f\|_\infty \leq 1} \sqrt{\mathbf{E}\,(\mu f - \nu f)^2}.$$

It is easy to verify that $\|\cdot\|$ and $|||\cdot|||$ define two metrics in the space of probability measures. Both metrics yields numbers between 0 (if $\mu = \nu$) and 2 (if $\mu \perp \nu$). In fact, if $\mu \perp \nu$ then there exists $A \in \mathcal{E}$ such that $\mu(A) = 0$ and $\nu(A) = 1$, and choosing $f = \mathbf{1}_A - \mathbf{1}_{A^c}$, where A^c is the complement of A, we have $|\mu f - \nu f| = 2$. Note that if μ and ν are not random, then we clearly have $|||\mu - \nu||| = \|\mu - \nu\|$.

We now present some results in the general setting of (possibly random) probability measures. These results hold both with respect to the metric $|||\cdot|||$ and with respect to the metric $\|\cdot\|$ (in the latter case, if the probably measures are random then these bounds hold for each realization of the randomness).

As we will be interested in conditional distributions, we need to understand how conditioning affects the distance between measures. Since conditioning introduces weights on measures (see Bayes formula, Theorem 2.7), we will need the following lemma.

Lemma 2.9 (Weighted measures)**.** *Let μ and ν be (possibly random) probability measures on a measurable space $(E, \mathcal{E})$, and let g be a real-valued $\mathcal{E}$-measurable function which is bounded away from zero and infinity, that is, $\inf_{x \in E} g(x) > 0$ and $\sup_{x \in E} g(x) < \infty$. Define*

$$\mu_g(A) := \frac{\int \mu(dx)\, g(x)\, \mathbf{1}_A(x)}{\int \mu(dx)\, g(x)}, \qquad \nu_g(A) := \frac{\int \nu(dx)\, g(x)\, \mathbf{1}_A(x)}{\int \nu(dx)\, g(x)}.$$

Then,

$$|||\mu_g - \nu_g||| \leq 2\, \frac{\sup_{x \in E} g(x)}{\inf_{x \in E} g(x)}\, |||\mu - \nu|||.$$

The same conclusion holds if the $|||\cdot|||$-norm is replaced by the $\|\cdot\|$-norm.

Proof. For any real-valued measurable function f we have

$$\frac{\mu(gf)}{\mu g} - \frac{\nu(gf)}{\nu g} = \frac{\mu(gf) - \nu(gf)}{\mu g} + \frac{\nu(gf)}{\mu g} - \frac{\nu(gf)}{\nu g}$$
$$= \frac{\|g\|_\infty}{\mu g}\left\{\mu\left(\frac{gf}{\|g\|_\infty}\right) - \nu\left(\frac{gf}{\|g\|_\infty}\right)\right\} + \frac{\nu(fg)}{\nu g}\frac{\|g\|_\infty}{\mu g}\left\{\nu\left(\frac{g}{\|g\|_\infty}\right) - \mu\left(\frac{g}{\|g\|_\infty}\right)\right\}.$$

If we assume that $\|f\|_\infty \leq 1$, then we have $\|\frac{gf}{\|g\|_\infty}\|_\infty \leq 1$ and $\nu(fg) \leq \nu(g)$, as g is positive. As $\|\frac{g}{\|g\|_\infty}\|_\infty \leq 1$ and $\mu g \geq \inf_x g(x)$, the proof follows immediately by using the triangle inequality for the metric $|||\cdot|||$ or for the metric $\|\cdot\|$. □

A collection of random variables $(X_n)_{n\geq 0}$ taking values in a measurable space $(E, \mathcal{E})$ is a *Markov chain* if there exists a transition kernel P from $(E, \mathcal{E})$ to $(E, \mathcal{E})$ such that for each $n \geq 1$ and each $A \in \mathcal{E}$ we have

$$\mathbf{P}(X_n \in A | X_0, \ldots, X_n) = P(X_{n-1}, A).$$

In this thesis we will be mostly interested in stochastic systems that can be described as Markov chains. Hence, we need to understand how the Markovian dynamics affects the distance between probability measures. The so-called *minorization condition* represents a strong condition that causes Markov chains to forget their initial condition at a geometric rate, as the following lemma shows.

Lemma 2.10 (Minorization condition for Markov chains)**.** *Let μ and ν be (possibly random) probability measures on a measurable space $(E, \mathcal{E})$ and let P be a transition kernel from $(E, \mathcal{E})$ to $(E, \mathcal{E})$. Then,*

$$|||\mu P - \nu P||| \leq |||\mu - \nu|||.$$

If there exist a probability measure ρ on $(E, \mathcal{E})$ and $\varepsilon > 0$ such that P satisfies the following minorization condition

$$P(x, A) \geq \varepsilon\rho(A) \qquad \text{for each } x \in E,\ A \in \mathcal{E},$$

then

$$|||\mu P - \nu P||| \leq (1 - \varepsilon)\,|||\mu - \nu|||.$$

The same conclusions hold if the $|||\cdot|||$-norm is replaced by the $\|\cdot\|$-norm.

Proof. The conditions $f \in \mathcal{E}$, $\|f\|_\infty \leq 1$ clearly imply $Pf \in \mathcal{E}$, $\|Pf\|_\infty \leq 1$. The first statement of the lemma follows immediately:

$$|||\mu P - \nu P||| = \sup_{f\in\mathcal{E}:\|f\|_\infty\leq 1} \sqrt{\mathbf{E}\,(\mu Pf - \nu Pf)^2} \leq |||\mu - \nu|||.$$

To prove the second statement, define

$$K(x, A) := \frac{P(x, A) - \varepsilon\rho(A)}{1 - \varepsilon}$$

for each $x \in E, A \in \mathcal{E}$. By the minorization condition it is easy to verify that K is a transition kernel. As

$$\mu P - \nu P = (1-\varepsilon)(\mu K - \nu K),$$

proceeding as above we get

$$|||\mu P - \nu P||| = (1-\varepsilon) \sup_{f \in \mathcal{E}: \|f\|_\infty \leq 1} \sqrt{\mathbf{E}\,(\mu K f - \nu K f)^2} \leq (1-\varepsilon)|||\mu - \nu|||.$$

The same argument holds with the $\|\cdot\|$-norm. □

Under the minorization condition the map $\mu \to \mu P$ is a strict contraction in the $|||\cdot|||$ norm. This implies that a Markov chains is *geometrically ergodic*: the difference of the law of the Markov chain started at two initial measures decays geometrically in time, namely,

$$|||\mu P^n - \nu P^n||| \leq (1-\varepsilon)^n\, |||\mu - \nu|||.$$

2.4 Distances between probability measures in high dimension

In this thesis we will be interested in the behavior of probability measures in high (possibly infinite) dimension. The canonical description of a high-dimensional random system is provided by specifying a probability measure ρ on a (possibly infinite) product space $E = \prod_{i \in I} E^i$: each site $i \in I$ represents a single degree of freedom, or dimension, of the model. When I is defined as the set of vertices of a graph, the measure ρ defines a graphical model or a random field. Models of this type are ubiquitous in statistical mechanics, combinatorics, computer science, statistics, and in many other areas of science and engineering.

Let ρ and $\tilde{\rho}$ be two such models that are defined on the same space E. We ask the following basic question: when is $\tilde{\rho}$ a good approximation of ρ? As briefly seen in the previous section, probability theory provides numerous methods to evaluate the difference between arbitrary probability measures. However, the high-dimensional setting brings some specific challenges: any approximation of practical utility in high dimension must yield error bounds that do not grow, or at least grow sufficiently slowly, in the model dimension $d = \operatorname{card} I$. We therefore seek quantitative methods that allow to establish dimension-free bounds on high-dimensional probability distributions.

The Dobrushin comparison theorem that we are about to introduce is a powerful (albeit blunt) tool to bound the total variation distance between marginals of high-dimensional probability measures ρ and $\tilde{\rho}$ in terms of their local conditional distributions. This method was developed by Dobrushin in [18, Theorem 3] in the context of statistical mechanics. Presently we introduce this tool in its simplified form, which is due to Föllmer [24] and has become standard textbook material, cf. [27, Theorem 8.20] and [45, Theorem V.2.2].[1] Despite the crucial importance that this

[1] Note that our definition of $\|\cdot\|_J$ differs by a factor 2 from that in [27].

theorem has for the results that will be developed in this thesis, we refer to Chapter 6 for its proof (see Section 6.3.1 in particular). In fact, one of the goal of Chapter 6 is precisely to develop a more general version of this comparison theorem.

Define the coordinate projections $X^i : x \mapsto x^i$ for $x \in E$ an $i \in I$. For any probability ρ on E, we fix a version $\rho^i_\cdot$ of the regular conditional probability

$$\rho^i_x(A) := \rho(X^i \in A | X^{I\setminus\{i\}} = x^{I\setminus\{i\}}).$$

We also define for $J \subseteq I$ the local total variation distance

$$\|\rho - \rho'\|_J := \sup_{f \in \mathcal{S}^J : |f| \le 1} |\rho(f) - \rho'(f)|,$$

where $\mathcal{S}^J$ is the class of measurable functions $f : E \to \bar{\mathbb{R}}$ such that $f(x) = f(z)$ whenever $x^J = z^J$. For $J = I$, we write $\|\rho - \rho'\|$ for simplicity.

Theorem 2.11 (Dobrushin comparison theorem)**.** *Let $\rho, \tilde{\rho}$ be probability measures on E. Define*

$$C_{ij} = \frac{1}{2} \sup_{x,z \in E : x^{I\setminus\{j\}} = z^{I\setminus\{j\}}} \|\rho^i_x - \rho^i_z\|, \qquad b_j = \sup_{x \in E} \|\rho^j_x - \tilde{\rho}^j_x\|.$$

Suppose that the Dobrushin condition holds:

$$\max_{i \in I} \sum_{j \in I} C_{ij} < 1.$$

Then the matrix sum $D := \sum_{n \ge 0} C^n$ is convergent, and we have for every $J \subseteq I$

$$\|\rho - \tilde{\rho}\|_J \le \sum_{i \in J} \sum_{j \in I} D_{ij} b_j.$$

The Dobrushin comparison theorem can be informally interpreted as follows. C_{ij} measures the degree to which a perturbation of site j directly affects site i under the distribution ρ. However, perturbing site j might also indirectly affect i: it could affect another site k which in turn affects i, etc. The aggregate effect of a perturbation of site j on site i is captured by the quantity D_{ij}. If D_{ij} decays exponentially in the distance $d(i,j)$ (which is a useful manifestation of the decay of correlations property that we will often encounter in this thesis), then Theorem 2.11 yields, for example, $\|\rho - \tilde{\rho}\|_i \lesssim \sum_j e^{-d(i,j)} b_j$, where b_j measures the local error at site j between ρ and $\tilde{\rho}$ (in terms of the conditional distributions $\rho^j_\cdot$ and $\tilde{\rho}^j_\cdot$).

In many applications it is natural to describe high-dimensional probability distributions in terms of local conditional probabilities of the form ρ^i_x. This is in essence a static picture, where we describe the behavior of each coordinate i given that the configuration of the remaining sites $I\setminus\{i\}$ is frozen. In models that possess dynamics, this description is not very natural. In this setting, each site $i \in I$ occurs at a given time $\tau(i)$, and its state is only determined by the configuration of sites $j \in I$ in the past and present $\tau(j) \le \tau(i)$, but not by the future. It is therefore interesting to note

that the original comparison theorem of Dobrushin [18] is actually more general than Theorem 2.11 in that it is applicable both in the static and dynamic settings. We presently state the one-sided counterpart to Theorem 2.11, and we refer to Chapter 6 for a more general version of this result and for its proof (see Section 6.3.3).

Assume that we are given a function $\tau : I \to \mathbb{Z}$ that assigns to each site $i \in I$ an integer index $\tau(i)$. Define

$$I_{\leq i} := \{j \in I : \tau(j) \leq i\}.$$

For any probability ρ on E, we fix a version $\gamma^i_\cdot$ of the regular conditional probability

$$\gamma^i_x(A) := \rho(X^i \in A | X^{I_{\leq \tau(i)} \setminus \{i\}} = x^{I_{\leq \tau(i)} \setminus \{i\}}).$$

We can now state the one-sided Dobrushin comparison theorem.

Theorem 2.12 (One-sided Dobrushin comparison theorem)**.** *Let $\rho, \tilde{\rho}$ be probability measures on E. Define*

$$C_{ij} = \frac{1}{2} \sup_{x,z \in E : x^{I \setminus \{j\}} = z^{I \setminus \{j\}}} \|\gamma^i_x - \gamma^i_z\|, \qquad b_j = \sup_{x \in E} \|\gamma^j_x - \tilde{\gamma}^j_x\|.$$

Suppose that the Dobrushin condition holds:

$$\max_{i \in I} \sum_{j \in I} C_{ij} < 1.$$

Then the matrix sum $D := \sum_{n \geq 0} C^n$ is convergent, and we have for every $J \subseteq I$

$$\|\rho - \tilde{\rho}\|_J \leq \sum_{i \in J} \sum_{j \in I} D_{ij} b_j.$$

Note that the one-sided comparison theorem can be interpreted as a generalization of Theorem 2.11 (just take τ to be a constant function). However, we stated two different theorems to stress the difference between the static and dynamic case. Both theorems will play a crucial role in this thesis.

In order to use these comparison theorems we must be able to bound the quantities C_{ij} and b_j. The elementary lemmas introduced in Section 2.3 will be used precisely for this purpose. We presently introduce a lemma that will be essential for bounding the matrix D coming from the comparison theorems. This result states that if C_{ij} decays exponentially in the distance between i and j at a sufficiently rapid rate, then D_{ij} will also decay exponentially in the distance between i and j. It is essentially a simple lemma about matrices.

Lemma 2.13. *Let I be a finite set and let m be a pseudometric on I. Let $C = (C_{ij})_{i,j \in I}$ be a matrix with nonnegative entries. Suppose that*

$$\max_{i \in I} \sum_{j \in I} e^{m(i,j)} C_{ij} \leq c < 1.$$

Then the matrix $D = \sum_{n\geq 0} C^n$ *satisfies*

$$\max_{i\in I} \sum_{j\in I} e^{m(i,j)} D_{ij} \leq \frac{1}{1-c}.$$

In particular, this implies that

$$\sum_{j\in J} D_{ij} \leq \frac{e^{-m(i,J)}}{1-c}$$

for every $J \subseteq I$.

Proof. Define for any matrix A with nonnegative entries the norm

$$\|A\|_m := \max_{i\in I} \sum_{j\in I} e^{m(i,j)} A_{ij}.$$

Using $m(i,j) \leq m(i,k) + m(k,j)$, we compute

$$\begin{aligned}
\|AB\|_m &= \max_{i\in I} \sum_{j\in I} e^{m(i,j)} \sum_{k\in I} A_{ik} B_{kj} \\
&\leq \max_{i\in I} \sum_{k\in I} e^{m(i,k)} A_{ik} \sum_{j\in I} e^{m(k,j)} B_{kj} \\
&\leq \|A\|_m \|B\|_m,
\end{aligned}$$

so $\|A\|_m$ is a matrix norm. Therefore,

$$\|D\|_m \leq \sum_{n\geq 0} \|C\|_m^n \leq \sum_{n\geq 0} c^n = \frac{1}{1-c}.$$

As

$$e^{m(i,J)} \sum_{j\in J} A_{ij} \leq \sum_{j\in J} e^{m(i,j)} A_{ij} \leq \|A\|_m,$$

the last statement of the lemma follows immediately. □

In the remainder of this section we present two simple results that are meant to illustrate the models that we will consider in this thesis.

Often we will state the general fact that "probability measures tend to be singular in high (or infinite) dimension." The following proposition exhibits a concrete manifestation of this general fact.

Proposition 2.14. *Let* $(E, \mathcal{E})$ *be a measurable space and let* μ *and* ν *be two probability measures on it. Define the following product measures on* $(E^{\mathbb{N}}, \mathcal{E}^{\mathbb{N}})$*:*

$$\mu^{\otimes} := \bigotimes_{n\in\mathbb{N}} \mu, \qquad \nu^{\otimes} := \bigotimes_{n\in\mathbb{N}} \nu.$$

If $\mu \neq \nu$ *then* $\mu^{\otimes} \perp \nu^{\otimes}$.

Proof. Let $(\Omega, \mathcal{H}, \mathbf{P})$ be a probability space, and let $(X_n)_{n\in\mathbb{N}}$ and $(Y_n)_{n\in\mathbb{N}}$ be two collections of i.i.d. random variables taking values in $(E, \mathcal{E})$, such that $X_n \sim \mu$ and $Y_n \sim \nu$ for each $n \in \mathbb{N}$. As $\mu \neq \nu$, there exists $A \in \mathcal{E}$ such that $\mu(A) \neq \nu(A)$. Define

$$B := \left\{ z = (z_1, z_2, \ldots) \in E^{\mathbb{N}} : \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} I_A(z_n) = \mu(A) \right\}.$$

By the law of large numbers we have

$$\mu^{\otimes}(B) = \mathbf{P}\left(\left\{ \omega \in \Omega : \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} I_A(X_n(\omega)) = \mu(A) \right\} \right) = 1,$$
$$\nu^{\otimes}(B) = \mathbf{P}\left(\left\{ \omega \in \Omega : \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} I_A(Y_n(\omega)) = \mu(A) \right\} \right) = 0.$$

□

Proposition 2.14 attests that unless two measures μ and ν are the same, their infinite products $\mu^{\otimes}$ and $\nu^{\otimes}$ are mutually singular. This example illustrates the fundamental reason why a *global* analysis of high-dimensional models is not suitable to properly describe these models. On the other hand, these models can be meaningfully interpreted by looking at *local* quantities. To see this, consider the case $\|\mu - \nu\| = \varepsilon \ll 1$. Then, for $J \subset I$ a telescoping argument easily gets $\|\mu^{\otimes} - \nu^{\otimes}\|_J \leq \varepsilon \operatorname{card} J$, whereas the global total variation bound yields $\|\mu^{\otimes} - \nu^{\otimes}\| = 2$. By bounding the local total variation distance over subsets of coordinates in terms of local quantities (the conditional distributions of each coordinate given all the others), the Dobrushin comparison theorem represents the key tool that will be used in this thesis to perform a local analysis in high-dimensional models.

While infinite-dimensional probability measures can be equivalent, they can differ *significantly* only on a finite number of coordinates, as the following example taken from [44, Chapter 9] illustrates.

Example 2.15. *For each $a \in \mathbb{R}$, let χ_a be the distribution of a Gaussian random variable in $\mathbb{R}$ with mean a and variance 1. Given two sequences $(a_n)_{n\in\mathbf{N}}$ and $(b_n)_{n\in\mathbf{N}}$, define the following product measures on $\mathbb{R}^{\mathbb{N}}$:*

$$\mu^{\otimes} := \bigotimes_n \chi_{a_n}, \qquad \nu^{\otimes} := \bigotimes_n \chi_{b_n}.$$

Then, $\mu^{\otimes} \sim \nu^{\otimes}$ if and only if $\sum_{n\in\mathbb{N}} (a_n - b_n)^2 < \infty$.

This example tells us that in order to be equivalent infinite-dimensional probability measures can only have finitely many degrees of freedom that can carry significant information. While such cases represent respectable infinite-dimensional models, for the sake of the results developed in this thesis we think of them as being *effectively finite-dimensional*. On the other hand, in this thesis we will be interested in models that are *genuinely infinite-dimensional*, in the sense that they are constituted by infinitely-many independent degrees of freedom.

2.5 Monte Carlo

Given a measurable space $(E, \mathcal{E})$ and a (possibly random) probability measure μ on it (perhaps known only up to a normalization factor), in this thesis we will be interested in the problem of approximating integrals of the form $\mu f = \int \mu(dx) f(x)$, for a suitable class of $\bar{\mathbb{R}}$-valued $\mathcal{E}$-measurable functions f. The Monte Carlo approach consists in approximating μf with the *sample mean* of f under μ. If μ is not random, this means:

$$\mu f = \mathbf{E}\, f(X) \approx \frac{1}{N} \sum_{i=1}^{N} f(X(i)),$$

where $X \sim \mu$, and $X(1), \ldots, X(N)$ are i.i.d. random variables (*samples*) with distribution μ, for a certain $N \geq 1$. On the other hand, if μ is random this means:

$$\mu f \equiv \mu f(Y) = \mathbf{E}(f(X)|Y) \approx \frac{1}{N} \sum_{i=1}^{N} f(X(i)),$$

where Y is the random variable through which the randomness in μ comes, as prescribed by Remark 2.5.

It is convenient to introduce the following sampling operator on probability measures (δ_x denotes the Dirac measure with mass located at $x \in E$).

Definition 2.16 (Sampling operator)**.** *Let μ be a (possibly random) probability measure on $(E, \mathcal{E})$. Define the* sampling operator S^N *as*

$$\mathsf{S}^N \mu := \frac{1}{N} \sum_{i=1}^{N} \delta_{X(i)}, \qquad X(1), \ldots, X(N) \text{ are i.i.d. samples} \sim \mu.$$

As $\mathsf{S}^N \mu$ is defined in terms of (possibly conditionally, cf. Remark 2.5) i.i.d. random variables, there are a lot of results to assess the accuracy of $(\mathsf{S}^N \mu) f$ as an estimate of μf. In particular, as N goes to infinity the strong Law of Large Numbers tells us that $(\mathsf{S}^N \mu) f$ converges almost surely to μf, while the Central Limit Theorem tells us that $\sqrt{N}\{(\mathsf{S}^N \mu) f - \mu f\}$ converges in distribution to a Gaussian with mean 0 and variance $\mu f^2 - (\mu f)^2$. Non-asymptotic results can also be easily obtained, such as bounds on tail probabilities $\mathbf{P}\{|(\mathsf{S}^N \mu) f - \mu f| \geq t\}$, for $t \geq 0$, and bounds on error moments $\mathbf{E}\,|(\mathsf{S}^N \mu) f - \mu f|^p$, for $p \geq 1$. We presently prove a result for the case $p = 2$, as this will be used repeatedly in this thesis. We refer to [8] for a systematic collection of these results.

Let us first recall the *bias/variance* decomposition of the mean square error, which is one of the most analytically tractable measure of the quality of an estimator:

$$\mathbf{E}\,((\mathsf{S}^N \mu) f - \mu f)^2 = \underbrace{\mathbf{E}\,((\mathsf{S}^N \mu) f - \mathbf{E}(\mathsf{S}^N \mu) f)^2}_{\text{variance}} + \underbrace{\left[\mathbf{E}\,(\mathsf{S}^N \mu) f - \mu f\right]^2}_{\text{bias}^2}.$$

Clearly, $(\mathsf{S}^N \mu) f$ is an *unbiased* estimator for each $N \geq 1$, as $\mathbf{E}\,(\mathsf{S}^N \mu) f = \mu f$. As for the variance of the estimator, we have the following lemma.

Lemma 2.17 (Monte Carlo variance)**.** *Let μ be a random probability measures on $(E, \mathcal{E})$, and let Y as in Remark 2.5. For each positive $\mathcal{E}$-measurable function f we have*

$$\mathbf{E}\left(((\mathsf{S}^N\mu)f - \mu f)^2 | Y\right) = \frac{1}{N}\left\{\mu(f^2) - (\mu f)^2\right\}.$$

As a consequence,

$$|||\mathsf{S}^N\mu - \mu||| \leq \frac{1}{\sqrt{N}}.$$

Proof. Note that

$$\begin{aligned}((\mathsf{S}^N\mu)f - \mu f)^2 = (\mu f)^2 - 2\frac{\mu f}{N}\sum_{i=1}^N f(X(i)) + \frac{1}{N^2}\sum_{i=1}^N (f(X(i)))^2 \\ + \frac{1}{N^2}\sum_{\substack{i,j\in\{1,\dots,N\}\\ i\neq j}} f(X(i))\, f(X(j)).\end{aligned}$$

By definition of the samples $X(1), \dots, X(N)$ (see also Remark 2.5) and by the properties of conditional expectations (recall that μf is σY-measurable), we have

$$\begin{aligned}\mathbf{E}\left(((\mathsf{S}^N\mu)f - \mu f)^2 | Y\right) &= (\mu f)^2 - 2\,(\mu f)^2 + \frac{1}{N}\,\mu(f^2) + \frac{N-1}{N}\,(\mu f)^2 \\ &= \frac{1}{N}\left\{\mu(f^2) - (\mu f)^2\right\},\end{aligned}$$

and the statement follows immediately. □

The Monte Carlo approximation scheme introduced above is practicable only when it is possible (and computationally convenient) to sample from the distribution μ itself, the so-called *target distribution*. More generally, there are situations where it is more convenient to sample from another distribution ν on $(E, \mathcal{E})$, which is then referred to as the *importance distribution* (or *proposal distribution*). The importance sampling paradigm is based on the idea that we can approximate μf using samples coming from ν. In fact, if $\mu \ll \nu$ then the Radon-Nikodym theorem (Theorem 2.6) yields

$$\mu f = \mathbf{E}\, f(X) = \mathbf{E}\, \tfrac{d\mu}{d\nu}(Z)\, f(Z) \approx \frac{1}{N}\sum_{i=1}^N \tfrac{d\mu}{d\nu}(Z(i))\, f(Z(i)),$$

where $X \sim \mu$, $Z \sim \nu$, and $Z(1), \dots, Z(N)$ are i.i.d. samples with distribution ν.

More generally, in this thesis we will deal with situations where the target distribution μ, or the instrumental distribution ν, or both, are only known up to a scalar factor. In this case also the Radon-Nikodym derivative $\frac{d\mu}{d\nu}$ is also known up to a constant factor. Nonetheless, the importance sampling paradigm can still be implemented by considering the following approximation where constant factors cancel out:

$$\mu f = \mathbf{E}\, f(X) = \mathbf{E}\, \tfrac{d\mu}{d\nu}(Z)\, f(Z) = \frac{\mathbf{E}\, \frac{d\mu}{d\nu}(Z)\, f(Z)}{\mathbf{E}\, \frac{d\mu}{d\nu}(Z)} \approx \frac{\sum_{i=1}^N \frac{d\mu}{d\nu}(Z(i))\, f(Z(i))}{\sum_{i=1}^N \frac{d\mu}{d\nu}(Z(i))},$$

where $X \sim \mu$, $Z \sim \nu$, $Z(1), \dots, Z(N)$ are i.i.d. samples with distribution ν, and we have used that $\mathbf{E}\, \frac{d\mu}{d\nu}(Z) = \mu(E) = 1$. The self-normalized importance sampling paradigm will be used in Chapter 3 to describe the basic algorithms upon which much of the work in this thesis is based. For this reason, we introduce a sampling operator also for this case.

Definition 2.18 (Self-normalized importance sampling operator)**.** *Let μ, ν be (possibly random) probability measures on $(E, \mathcal{E})$ such that $\mu \ll \nu$. Define the* self-normalized importance sampling operator S_ν^N *as*

$$\mathsf{S}_\nu^N \mu := \sum_{i=1}^{N} W(i)\, \delta_{Z(i)}, \qquad Z(1), \dots, Z(N) \text{ are i.i.d. samples} \sim \nu,$$

where the weights $W(1), \dots, W(N)$ are defined as

$$W(i) := \frac{\frac{d\mu}{d\nu}(Z(i))}{\sum_{\ell=1}^{N} \frac{d\mu}{d\nu}(Z(\ell))}.$$

Clearly, we have $\mathsf{S}_\mu^N \mu = \mathsf{S}^N \mu$. Consistency and asymptotic normality are easy to prove, and now $\sqrt{N}\,\{(\mathsf{S}^N \mu) f - \mu f\}$ converges in distribution to a Gaussian with mean 0 and variance $\mathbf{E}\,(\frac{d\mu}{d\nu}(Z)\,(f(Z) - \mu f))^2$, $Z \sim \nu$. The self-normalize estimate $(\mathsf{S}_\nu^N \mu) f$ is biased for any fixed value of N, and establishing non-asymptotic results is not as straightforward as for the ordinary Monte Carlo approximation. We refer again to [8] for details.

Chapter 3

Classical nonlinear filtering and particle filters

This chapter provides an overview of the classical theory of nonlinear filtering and sequential Monte Carlo algorithms known as particle filters. Emphasis is given to the stability property of the filter distribution, which is the key to establish time-uniform error bounds for particle filters. The treatment revolves around the curse of dimensionality phenomenon, and the coverage is instrumental to the content that will be developed in Chapter 4 and Chapter 5. The presentation is inspired by [55] and [8].

3.1 Hidden Markov models and nonlinear filter

Let $(\mathbb{X}, \mathcal{X})$ and $(\mathbb{Y}, \mathcal{Y})$ be two Polish spaces. We define a *hidden Markov model* as a $(\mathbb{X} \times \mathbb{Y}, \mathcal{X} \otimes \mathcal{Y})$-measurable Markov chain $(X_n, Y_n)_{n \geq 0}$ whose transition probability kernel K can be factored as

$$K f(x, y) = \int p(x, x')\, g(x', y')\, \psi(dx')\, \varphi(dy')\, f(x', y'),$$

for each $x \in \mathbb{X}, y \in \mathbb{Y}$ and each $\mathcal{X} \otimes \mathcal{Y}$-measurable function f. Thus, $(X_n)_{n \geq 0}$ is itself a Markov chain in $(\mathbb{X}, \mathcal{X})$ with transition density $p : \mathbb{X} \times \mathbb{X} \to \bar{\mathbb{R}}_+$ with respect to a given reference measure ψ, while $(Y_n)_{n \geq 0}$ are random variables in $(\mathbb{Y}, \mathcal{Y})$ that are conditionally independent given $(X_n)_{n \geq 0}$ with transition density $g : \mathbb{X} \times \mathbb{Y} \to \bar{\mathbb{R}}_+$ with respect to a reference measure φ. This dependency structure is illustrated in Figure 3.1. We interpret $(X_n)_{n \geq 0}$ as an underlying dynamical process—the signal—that is not directly observable, while the observable process $(Y_n)_{n \geq 0}$ consists of partial and noisy observations of $(X_n)_{n \geq 0}$. The hidden Markov model setting is convenient mathematically and is ubiquitous in practice as a model of noisy observations of random dynamics.

In the following we will assume that the process $(X_n, Y_n)_{n \geq 0}$ is realized on its canonical probability space, and denote for any probability measure μ on $(\mathbb{X}, \mathcal{X})$ by $\mathbf{P}^\mu$ the probability measure under which $(X_n, Y_n)_{n \geq 0}$ is a hidden Markov model with

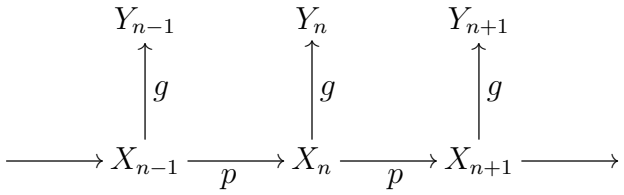

Figure 3.1: Dependency graph of a hidden Markov model.

transition probability P as above and with initial condition $X_0 \sim \mu$ (if we simply write $\mathbf{P}$, then it means that any choice of the initial measure would yield equivalent results for the argument being considered). For $x \in \mathbb{X}$, we write for simplicity $\mathbf{P}^x := \mathbf{P}^{\delta_x}$. As the process $(X_n)_{n\geq 0}$ is unobservable, a central problem in this setting is to track the unobserved state X_n given the observation history $Y_1, \ldots, Y_n$: that is, we aim to compute the *nonlinear filter*

$$\pi_n^\mu := \mathbf{P}^\mu(X_n \in \cdot\,|Y_1, \ldots, Y_n).$$

Filtering—the computation of the conditional distributions of a hidden Markov process given observed data—is a problem that arises in a wide array of applications in science and engineering, classically in the field of tracking, speech recognition, and finance. We refer to [8] for a rich list of applications.

Remark 3.1 (A matter of notation)**.** *To be precise, given our definition of conditional distributions (Definition 2.4), we should write $\pi_n^\mu(Y_{1:n}, \cdot\,)$ instead of π_n^μ. However, in what follows we only use the kernel notation $\pi_n^\mu(y_{1:n}, dx)$ to emphasize the dependence of the filter on a particular sequence of observations $Y_{1:n} = y_{1:n}$. Hence, we interpret π_n^μ as a random measure whose randomness is (implicitly) provided by the observations $Y_1, \ldots, Y_n$.*

Being a conditional distribution, the filter yields least mean square estimates, and for this reason it is often referred to as the *optimal* filter.

Lemma 3.2 (Optimality of the filter)**.** *Fix $n \geq 0$. Let f be a measurable function such that $\mathbf{E}^\mu f(X_n)^2 < \infty$. Then,*

$$\pi_n^\mu f = \arg\min_h \mathbf{E}^\mu \left(f(X_n) - h(Y_{1:n})\right)^2,$$

where the minimization is over measurable functions h.

Proof. It follows immediately from Lemma 2.2, choosing $X = f(X_n)$ and $Y = (Y_1, \ldots, Y_n)$. □

If the conditional distribution π_n can be computed, it yields not only a least mean square estimate of the unobserved state X_n, but also a complete representation of the uncertainty in this estimate.

An important property of the filter is that it can be computed recursively, which follows immediately from Bayes formula (Lemma 2.8).

Lemma 3.3 (Filter recursion)**.** *The filter distribution π_n^μ can be computed recursively according to*

$$\pi_n^\mu f = \frac{\int \pi_{n-1}^\mu(dx)\, p(x,x')\, \psi(dx')\, g(x', Y_n)\, f(x')}{\int \pi_{n-1}^\mu(dx)\, p(x,x')\, \psi(dx')\, g(x', Y_n)}.$$

with the initial condition $\pi_0^\mu = \mu$.

Proof. Fix $n \geq 0$. By construction $X_{0:n}$ has distribution ρ given by

$$\rho(dx_{0:n}) := \mathbf{P}^\mu(X_{0:n} \in dx_{0:n}) = \mu(dx_0)\, p(x_0, x_1)\psi(dx_1) \cdots p(x_{n-1}, x_n)\psi(dx_n).$$

Define the provability measure λ on $(\mathbb{Y}^n, \mathcal{Y}^n)$ as

$$\lambda(dy_{1:n}) := \varphi(dy_1) \cdots \varphi(dy_n),$$

and define the positive function γ as

$$\gamma(x_{1:n}, y_{1:n}) := g(x_1, y_1) \cdots g(x_n, y_n).$$

By construction we have

$$\mathbf{E}^\mu\, f(X_{0:n}, Y_{1:n}) = \int \rho(dx_{0:n})\, \lambda(dy_{1:n})\, \gamma(x_{1:n}, y_{1:n})\, f(x_{0:n}, y_{1:n})$$

for each positive measurable function f. By Bayes formula (Lemma 2.8) we have that the conditional distribution of $X_{0:n}$ given $Y_{1:n}$ is given by the probability kernel P defined as

$$Pf(Y_{1:n}) = \int \mathbf{P}^\mu(X_{0:n} \in dx_{0:n} | Y_{1:n})\, f(x_{0:n}) = \frac{\int \rho(dx_{0:n})\, \gamma(x_{0:n}, Y_{1:n})\, f(x_{0:n})}{\int \rho(dx_{0:n})\, \gamma(x_{0:n}, Y_{1:n})}.$$

It is immediately verified that

$$\begin{aligned}\pi_n^\mu f &= \int P(Y_{1:n}, dx_{0:n})\, f(x_n) = \frac{\int \rho(dx_{0:n})\, \gamma(x_{0:n}, Y_{1:n})\, f(x_n)}{\int \rho(dx_{0:n})\, \gamma(x_{0:n}, Y_{1:n})} \\ &= \frac{\int \pi_{n-1}^\mu(dx)\, p(x,x')\, \psi(dx')\, g(x', Y_n)\, f(x')}{\int \pi_{n-1}^\mu(dx)\, p(x,x')\, \psi(dx')\, g(x', Y_n)}.\end{aligned}$$

□

The recursive structure of the nonlinear filter is of central importance, as it allows the filter to be computed on-line over a long time horizon. Nonetheless, the recursion is still at the level of probability measures, and in general no finite-dimensional sufficient statistics exist. Important exceptions are two special cases: linear Gaussian models (which give rise to the celebrated *Kalman filter*) and models with a (small) finite state space, cf. [8]. However, most complex models do not fall into these very limited categories. Therefore, the practical implementation of nonlinear filters typically proceeds by sequential Monte Carlo approximations known as *particle filters*. We refer to [19] for a survey on these methods. In the present context we limit ourself to describe, in their basic formulations, the main two algorithms that have been considered in the filtering literature. We present these algorithms in the light of the *curse of dimensionality* phenomenon that affects both of them, which will be instrumental for the material developed in Chapter 4 and Chapter 5.

3.2 Sequential importance sampling

One of the first Monte Carlo algorithm that was used to approximate the filter distribution is the *sequential importance sampling* (SIS) particle filter. The introduction of this algorithm can be traced back to the pioneering work of Handschin and Mayne in 1969 [30]. The idea behind the SIS algorithm is to apply the self-normalized importance sampling paradigm introduced in Section 2.5 to approximate the so-called *smoothing* distribution $\mathbf{P}^\mu(X_{0:n} \in \cdot\,|Y_{1:n})$, and then compute the marginal at time n to approximate the filter $\pi_n^\mu = \mathbf{P}^\mu(X_n \in \cdot\,|Y_{1:n})$.

To see how the SIS works, fix $n \geq 1$ and assume that we are given the observations $Y_1, \dots, Y_n$. Our goal is to approximate integrals with respect to the (random) measure π_n^μ. From the proof of Lemma 3.3 we know that the conditional distribution of $X_{0:n}$ given $Y_{1:n}$ is given by the kernel P defined as

$$P(Y_{1:n}, dx_{0:n}) := \mathbf{P}^\mu(X_{0:n} \in dx_{0:n}|Y_{1:n}) = \frac{1}{Z}\int \rho(dx_{0:n})\, g(x_1, Y_1)\cdots g(x_n, Y_n),$$

where

$$\rho(dx_{0:n}) := \mathbf{P}^\mu(X_{0:n} \in dx_{0:n}) = \mu(dx_0)\, p(x_0, x_1)\psi(dx_1)\cdots p(x_{n-1}, x_n)\psi(dx_n)$$

and

$$Z := \int \rho(dx_{0:n})\, g(x_1, Y_1)\cdots g(x_n, Y_n). \tag{3.1}$$

At first sight, we might think of using straightforwardly the Monte Carlo approximation (recall the definition of the sampling operator S^N, Definition 2.16)

$$\pi_n^\mu f(Y_{1:n}) \approx \int (\mathsf{S}^N P_{Y_{1:n}})(dx_{0:n})\, f(x_n) = \frac{1}{N}\sum_{i=1}^N f(X_n(i)),$$

where, for each $i \in \{1, \dots, N\}$, $X(i) := (X_0(i), \dots, X_n(i))$ is an independent sample from the distribution $P_{Y_{1:n}}$ (conditionally independent given $Y_1, \dots, Y_n$, see Remark 2.5). Of course, the problem with this approach is that in general we do not know how to sample from $P_{Y_{1:n}}$. However, by construction it is usually easy to sample from the signal Markov chain $(X_n)_{n\geq 0}$. This is the case, for instance, if the signal is modeled as a recursion

$$X_n = h(X_{n-1}, \xi_n), \qquad n \geq 1,$$

where $(\xi_n)_{n\geq 1}$ are i.i.d. random variables having a distribution that can be efficiently sampled (for instance, the uniform distribution or the Gaussian distribution), and h is a non-random function that we know pointwise. In this case, in fact, we can sample $X_n \sim p(x_{n-1}, \cdot\,)\psi$ by sampling ξ_n first, and then computing $X_n = h(x_{n-1}, \xi_n)$.
This fact suggests to use importance sampling choosing ρ as importance distribution and $P_{Y_{1:n}}$ as target distribution. The Radon-Nikodym derivative reads

$$\frac{dP_{Y_{1:n}}}{d\rho}(x_{0:n}) = \frac{1}{Z}\, g(x_1, Y_1)\cdots g(x_n, Y_n).$$

Since the normalization constant Z is not easy to compute (else, again, computing the filter distribution would not be a problem in the first place), then we apply the self-nomalized importance sampling operator S_ρ^N (Definition 2.18) to get

$$\pi_n^\mu f(Y_{1:n}) \approx \int (\mathsf{S}_\rho^N P_{Y_{1:n}})(dx_{0:n})\, f(x_n) = \sum_{i=1}^N W_n(i)\, f(X_n(i)),$$

where for each $i \in \{1, \dots, N\}$ we have

$$W_n(i) := \frac{g(X_1(i), Y_1) \cdots g(X_n(i), Y_n)}{\sum_{\ell=1}^N g(X_1(\ell), Y_1) \cdots g(X_n(\ell), Y_n)}$$

and $X(i) := (X_0(i), \dots, X_n(i))$ is an independent sample from the distribution ρ. Note that the weights $W_n(1), \dots, W_n(N)$ are positive and they sum to 1, and they depend on the (random) observation sequence $Y_1, \dots, Y_n$. So, the SIS particle filter approximation at time n is given by

$$\bar{\pi}_n^\mu(dx_n) := \int_{x_{0:n-1} \in \mathbb{X}^n} (\mathsf{S}_\rho^N P_{Y_{1:n}})(dx_{0:n}) = \sum_{i=1}^N W_n(i)\, \delta_{X_n(i)}(dx_n).$$

A key observation is that the weights can be computed recursively, namely,

$$W_n(i) \propto W_{n-1}(i)\, g(X_n(i), Y_n), \qquad W_0(i) = 1/N, \tag{3.2}$$

where the proportionality is up to the normalization factor so that $\sum_{i=1}^N W_n(i) = 1$. This fact suggests that the SIS particle filter can be implemented in an on-line fashion, as described in Figure 3.2. Figure 3.3 illustrates a typical iteration of the algorithm.

Algorithm 1: SIS particle filter

Data: Fix $n, N \geq 1$. Let the observations $Y_1, \dots, Y_n$ be given.
Sample $X_0(i)$, $i = 1, \dots, N$ from the initial distribution μ;
Set $W_0(i) = 1/N$, $i = 1, \dots, N$;
for $k = 1, \dots, n$ **do**
  Sample i.i.d. $X_k(i) \sim p(X_{k-1}(i), \cdot)\, d\psi$, $i = 1, \dots, N$;
  Compute $W_k(i) = W_{k-1}(i)\, g(X_k(i), Y_k) / \sum_{\ell=1}^N W_{k-1}(\ell)\, g(X_k(\ell), Y_k)$, $i = 1, \dots, N$;
Let $\bar{\pi}_n^\mu = \sum_{i=1}^N W_n(i)\, \delta_{X_n(i)}$;
Compute the approximate filter $\pi_n^\mu f \approx \bar{\pi}_n^\mu f$.

Figure 3.2: The classical sequential importance sampling (SIS) particle filter.

For any fixed time $n \geq 1$, the quality of the estimates obtained by the SIS particle filter as a function of the number of particles N can be easily assessed by the general theory on self-normalized importance sampling, see Section 2.5. In particular, the

SIS particle filter does indeed approximate the exact nonlinear filter as N goes to infinity with the typical Monte Carlo $1/\sqrt{N}$-rate for the mean-square error, namely,

$$\sqrt{\mathbf{E}\,(\pi_n^\mu f - \bar{\pi}_n^\mu f)^2} \leq \frac{C_n}{\sqrt{N}},$$

where C_n is a constant that depends on time n.

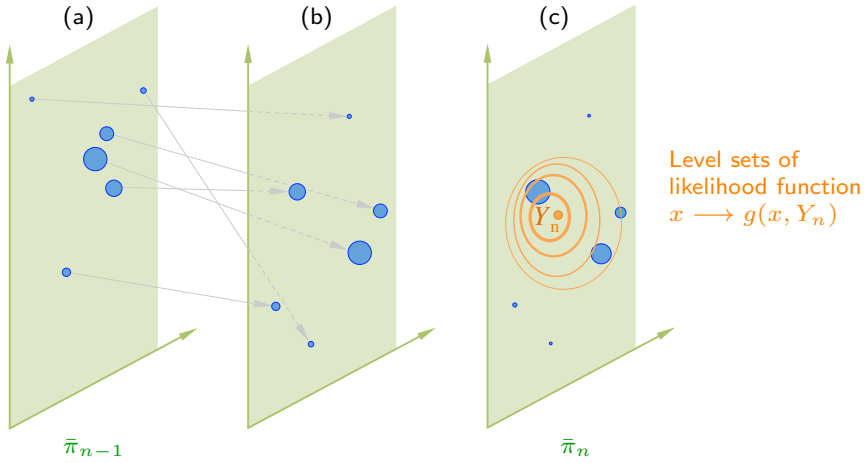

Figure 3.3: Representation of a single iteration of the SIS particle filter in the case when the state and observation spaces are $\mathbb{X} = \mathbb{Y} = \mathbb{R}_+^2$, and when there are $N = 6$ particles considered by the algorithm. Each particle is represented by a blue ball, whose size is proportional to the weight of the particle. (a) Representation of $\bar{\pi}_{n-1}$. (b) Particles are propagated forward using the underlying dynamics. (c) Particles are reweighed according to the likelihood of the new observation at time n (whose level sets are drawn in orange) yielding $\bar{\pi}_n$, following the multiplicative weight recursion (3.2).

3.2.1 Sample degeneracy with time

The SIS algorithm is a sequential implementation of the general importance sampling paradigm ("sequential" in the sense that there is no need of regenerating the populations of samples from scratch at the arrival of new observations). It turns out that importance sampling is usually very inefficient in high-dimensional models, so that the SIS particle filter performs poorly as time increases. The issue comes from the fact that importance sampling employs a finite number of samples from $\mathbf{P}^\mu(X_{0:n} \in \cdot\,)$ to approximate the target distribution $\mathbf{P}^\mu(X_{0:n} \in \cdot\,|Y_{1:n})$, and the approximation does not work well if the two distributions are too far apart, which is what happens if time n is large. In practice the SIS algorithm fails because the distribution of the weights $W_n(1), \ldots, W_n(N)$ degenerates as time n increases, and essentially only one particle is left with a non-zero weight after a few time steps (recall that at each time step the the weights sum to 1 by construction). This phenomenon is known as *collapse* or *sample/weight degeneracy*. The following example clarifies this issue.

Example 3.4 (Weight degeneracy of SIS with time)**.** *In the framework introduced in Section 3.1, consider the hidden Markov model where $(X_n)_{n\geq 0}$ is a symmetric random walk in $\mathbb{Z}^2$ with $X_0 = x \in \mathbb{Z}^2$, and for each $n \geq 1$ we have $Y_n = X_n + \varepsilon\eta_n$, where $\varepsilon \in \mathbb{R}_+$ and $(\eta_n)_{n\geq 0}$ is a collection of i.i.d. random variables having the standard Gaussian distribution in $\mathbb{R}^2$ (zero mean and identity covariance matrix). If the* signal-to-noise ratio *is high, that is, if ε is very close to 0, then we expect the smoothing distribution $\mathbf{P}^x(X_{0:n} \in \cdot\,|Y_{1:n})$ to be very concentrated around $X_{0:n}$, the true location of the path of the signal up to time n. However, if we sample N particles from the distribution $\mathbf{P}^x(X_{0:n} \in \cdot\,)$, where each particle represents a path of n steps of the symmetric random walk, then only a fraction of the particles will be close to any given trajectory in $\mathbb{Z}^2$, and the problem clearly gets worse as time increases.*

The phenomenon of weight degeneracy of the SIS algorithm with time has been analyzed in various settings. The following example (adapted from Example 7.3.1 in [8]) analyzes the poor performance of the SIS algorithm asymptotically (in the limit $N \to \infty$) as time increases.

Example 3.5 (Exponential growth of the SIS asymptotic variance with time)**.** *In the general framework introduced in Section 3.1, consider the hidden Markov model where $(X_n)_{n\geq 0}$ is a product of i.i.d. random variables with distribution μ (that is, $p(x, \cdot)\psi = \mu$ for each $x \in \mathbb{X}$). Then, for each time $n \geq 1$ we have*

$$N^{1/2}(\bar{\pi}_n^\mu f - \pi_n^\mu f) \xrightarrow{\text{in distribution}} \text{Gaussian}\left(0, \sigma_n^2(f)\right) \qquad \text{as } N \to \infty,$$

with

$$\sigma_n^2(f) := c(f)\,\gamma^{n-1},$$

where $c(f)$ and γ are constants that do not depend on n, $c(f) > 0$ as long as f is not a constant, and $\gamma > 1$ as long as the observation density g is different from 1.

First of all, as $(X_n)_{n\geq 0}$ is a collection of i.i.d. random variables with distribution μ, it follows that also $(Y_n)_{n\geq 1}$ is a collection of i.i.d. random variables with distribution

$$\mathbf{P}(Y_1 \in A) = \int \mu(dx)\, g(x, y)\, \varphi(dy)\, \mathbf{1}_A(y).$$

For each $x \in \mathbb{X}, y \in \mathbb{Y}$ define

$$\bar{g}(x, y) := \frac{g(x, y)}{\int \mu(dx)\, g(x, y)}.$$

Then, for each $N \geq 1, n \geq 1$ we have

$$N^{1/2}(\bar{\pi}_n^\mu f - \pi_n^\mu f) = \frac{N^{-1/2} \sum_{i=1}^N (f(X_n(i)) - \pi_n^\mu f) \prod_{k=1}^n \bar{g}(X_k(i), Y_k)}{N^{-1} \sum_{i=1}^N \prod_{k=1}^n \bar{g}(X_k(i), Y_k)}, \tag{3.3}$$

where $(X_k(i))$, $i \in \{1, \ldots, N\}$, $k \in \{1, \ldots, n\}$, is a collection of i.i.d. random variables with distribution μ, conditionally independent given $Y_1, \ldots, Y_n$. By independence, for each i and k we have

$$\mathbf{E}\,\bar{g}(X_k(i), Y_k) = \mathbf{E}\,\mathbf{E}(\bar{g}(X_k(i), Y_k)|Y_k) = \mathbf{E} \int \mu(dx)\bar{g}(x, Y_k) = 1,$$

and the strong Law of Large Numbers yields, as $N \to \infty$,

$$N^{-1} \sum_{i=1}^{N} \prod_{k=1}^{n} \bar{g}(X_k(i), Y_k) \xrightarrow{\text{almost surely}} \mathbf{E} \prod_{k=1}^{n} \bar{g}(X_k(1), Y_k) = \prod_{k=1}^{n} \mathbf{E}\, \bar{g}(X_k(1), Y_k) = 1$$

for the denominator in (3.3). *On the other hand, as*

$$\pi_n^\mu f = \int \mu(dx)\, \bar{g}(x, Y_n)\, f(x),$$

by independence it is immediately verified that

$$\mathbf{E}\, (f(X_n(1)) - \pi_n^\mu f) \prod_{k=1}^{n} \bar{g}(X_k(1), Y_k) = 0$$

and

$$\sigma_n^2(f) := \mathbf{E} \left((f(X_n(1)) - \pi_n^\mu f) \prod_{k=1}^{n} \bar{g}(X_k(1), Y_k) \right)^2 = c(f)\, \gamma^{n-1},$$

where

$$c(f) := \mathbf{E} \int \mu(dx)\, (f(x) - \pi_1^\mu f)^2\, \bar{g}(x, Y_1)^2,$$
$$\gamma := \mathbf{E} \int \mu(dx)\, \bar{g}(x, Y_1)^2.$$

The Central Limit Theorem yields that the numerator in (3.3) *converges in distribution as* $N \to \infty$ *to a Gaussian distribution with mean* 0 *and variance* $\sigma_n^2(f)$. *Therefore, it is immediate that also* (3.3) *converges in distribution to the same Gaussian distribution. Applying Jensen's inequality twice we get*

$$1 = \left(\mathbf{E} \int \mu(dx)\, \bar{g}(x, Y_1) \right)^2 \leq \mathbf{E} \left(\int \mu(dx)\, \bar{g}(x, Y_1) \right)^2 \leq \mathbf{E} \int \mu(dx)\, \bar{g}(x, Y_1)^2 = \gamma.$$

Thus, the asymptotic variance of the SIS algorithm increases exponentially with time as long as g *is different from* 1.

The analysis in Example 3.5 can be extended to more general models. However, even for linear Gaussian models where computations can be carried out explicitly, the analysis becomes much more involved (we refer to [8] and references therein). In practice, weight degeneracy is a major limitation that has render the SIS particle filter largely useless in many applications where one is interested in tracking the underlying state reliably for more than a few time steps.

In the next section we show that a modification of the sampling scheme considered so far can produce samples that have a closer distribution to the filter π_n^μ. This yields a new algorithm that can overcome the degeneracy of the weights with time.

Remark 3.6 (Importance sampling)**.** *In the literature (see [19] for instance) the term "sequential importance sampling" is generally used to indicate a more general algorithm than the one we just described. This term is used in the case where the importance distribution being used in the importance sampling paradigm corresponds to the law ρ' of a given (possibly time-inhomogeneous) Markov chain $(Z_n)_{n\geq 0}$, which can differ from the law ρ of the signal Markov chain $(X_n)_{n\geq 0}$. The idea is to choose an importance distribution that is as close as possible to the target distribution $\mathbf{P}^\mu(X_{0:n} \in \cdot\,|Y_{1:n})$, so to improve the performance of the algorithm and possibly alleviate weights degeneracy with time. Presently, we limit ourself to describe this more general version of the SIS algorithm, and we refer to the discussion developed in Section 3.3.3 to understand why importance sampling can not tackle the curse of dimensionality at a fundamental level.*

To make the point, fix $n \geq 1$, define

$$\rho'(dx_{0:n}) := \mu(dx_0)\, q_1(x_0, x_1)\psi(dx_1) \cdots q_n(x_{n-1}, x_n)\psi(dx_n),$$

and assume that for each $k \in \{1, \ldots, n\}$

$$(x, A) \in (\mathbb{X}, \mathcal{X}) \longrightarrow \int q_k(x, x')\psi(dx')\, \mathbf{1}_A(x')$$

is a given transition kernel so that $p(x, \cdot\,)\psi \ll q_k(x, \cdot\,)\psi$ for each $x \in \mathbb{X}$. Then, the Radon-Nikodym derivative reads

$$\frac{dP_{Y_{1:n}}}{d\rho'}(x_{0:n}) = \frac{1}{Z}\, \frac{p(x_0, x_1)\, g(x_1, Y_1)}{q_1(x_0, x_1)} \cdots \frac{p(x_{n-1}, x_n)\, g(x_n, Y_n)}{q_n(x_{n-1}, x_n)},$$

where Z is defined in (3.1). In this case the self-normalized importance sampling paradigm yields

$$\pi_n^\mu f(Y_{1:n}) \approx \int (\mathsf{S}_{\rho'}^N P_{Y_{1:n}})(dx_{0:n})\, f(x_n) = \sum_{i=1}^N W_n(i)\, f(Z_n(i)),$$

where for each $i \in \{1, \ldots, N\}$ the weight recursion now reads

$$W_n(i) \propto W_{n-1}(i)\, \frac{p(Z_{n-1}(i), Z_n(i))\, g(Z_n(i), Y_n)}{q_n(Z_{n-1}(i), Z_n(i))}, \qquad W_0(i) = 1/N,$$

(the proportionality is always up to the normalization factor so that $\sum_{i=1}^N W_n(i) = 1$), and each $Z(i) := (Z_0(i), \ldots, Z_n(i))$ is an independent sample from the distribution ρ', conditionally independent given $Y_1, \ldots, Y_n$ (see Remark 2.5). Clearly, if we choose $q_1, \ldots, q_n$ as

$$q_k(x, x')\psi(dx') := \mathbf{P}(X_k \in dx' | X_{k-1} = x) = p(x, x')\psi(dx'),$$

then we recover the SIS algorithm introduced in the main text. Another popular choice in the literature is given by

$$q_k^\star(x, x')\psi(dx') := \mathbf{P}(X_k \in dx' | X_{k-1} = x, Y_{1:k}) = \frac{p(x, x')\, g(x', Y_k)}{\int p(x, x')\, g(x', Y_k)\, \psi(dx')}\, \psi(dx'),$$

which yields the following weight recursion

$$W_k^\star(i) \propto W_{k-1}^\star(i) \int p(Z_{k-1}(i), x')\, g(x', Y_k)\, \psi(dx'), \qquad W_0(i) = 1/N.$$

The distribution $\rho'^\star$ obtained with this choice is the so-called optimal distribution. *In this context the adjective "optimal" refers to the fact that the conditional variance of the weights at each time step (given all the samples already generated by the algorithm) is zero, namely,*

$$\mathbf{Var}(W_n^\star(i) \,|\, Z_k(j), k \in \{1, \dots, n-1\}, j \in \{1, \dots, N\}) = 0,$$

as $W_n^\star(i)$ does not depend on $Z_n(i)$, $i \in \{1, \dots, N\}$.

3.3 Sequential importance resampling

One of the key property of the filter distribution is that it can be computed recursively: in order to compute π_n^μ we only need to know π_{n-1}^μ and Y_n (Lemma 3.3). Despite the fact that the SIS algorithm has an iterative implementation (Figure 3.2), the way we derived this algorithm does not capture the recursive structure of the filter, as the importance sampling paradigm was applied to the entire smoothing distribution $\mathbf{P}^\mu(X_{0:n} \in \cdot\,|Y_{1:n})$, for a fixed time n.

It seems natural to seek for a Monte Carlo approximation that can match the recursive nature of the filter. The most popular algorithm of this type is the *sequential importance resampling* (SIR) particle filter (also known as *bootstrap particle filter*) introduced in 1993 by Gordon, Salmond and Smith in 1993 [28], which simply inserts a sampling step in the filter recursion. To define this algorithm, let us rewrite the Bayes recursion as follows:

$$\pi_0^\mu = \mu, \qquad\qquad \pi_n^\mu = \mathsf{F}_n \pi_{n-1}^\mu \quad (n \geq 1),$$

where

$$(\mathsf{F}_n \rho) f := \frac{\int \rho(dx)\, p(x, x')\, \psi(dx')\, g(x', Y_n)\, f(x')}{\int \rho(dx)\, p(x, x')\, \psi(dx')\, g(x', Y_n)}.$$

It is instructive to write the recursion $\mathsf{F}_n := \mathsf{C}_n \mathsf{P}$ in two steps:

$$\pi_{n-1}^\mu \xrightarrow{\text{prediction}} \mathsf{P}\pi_{n-1}^\mu \xrightarrow{\text{correction}} \pi_n^\mu = \mathsf{C}_n \mathsf{P} \pi_{n-1}^\mu,$$

where

$$(\mathsf{P}\rho) f := \int \rho(dx)\, p(x, x')\, \psi(dx')\, f(x'),$$
$$(\mathsf{C}_n \rho) f := \frac{\int \rho(dx)\, g(x, Y_n)\, f(x)}{\int \rho(dx)\, g(x, Y_n)}.$$

In the prediction step, the filter π_{n-1}^μ is propagated forward using the dynamics of the underlying unobserved process $(X_n)_{n \geq 0}$ to compute the predictive distribution

$\mathbf{P}^\mu(X_n \in \cdot\,|Y_1,\dots,Y_{n-1})$. Then, in the correction step the predictive distribution is conditioned on the new observation Y_n to obtain the filter π_n^μ.

The SIR algorithm approximates π_n^μ by the empirical distribution $\hat\pi_n^\mu$ computed by the recursion

$$\hat\pi_0^\mu := \mu, \qquad\qquad \hat\pi_n^\mu := \hat{\mathsf{F}}_n \hat\pi_{n-1}^\mu \quad (n \geq 1),$$

where $\hat{\mathsf{F}}_n := \mathsf{C}_n \mathsf{S}^N \mathsf{P}$ consists of three steps

$$\hat\pi_{n-1}^\mu \xrightarrow{\text{prediction}} \mathsf{P}\hat\pi_{n-1}^\mu \xrightarrow{\text{sampling}} \mathsf{S}^N\mathsf{P}\hat\pi_{n-1}^\mu \xrightarrow{\text{correction}} \hat\pi_n^\mu := \mathsf{C}_n\mathsf{S}^N\mathsf{P}\hat\pi_{n-1}^\mu.$$

Here $N \geq 1$ is the number of particles used in the algorithm, and S^N is the sampling operator defined in Definition 2.16.[1]

It is straightforward to check that if $Z \sim \rho$ and $Z' \sim P(Z, \cdot\,)$, then $Z' \sim \mathsf{P}\rho$. So, at each time step $n \geq 1$, in order to draw N independent samples from $\mathsf{P}\hat\pi_{n-1}^\mu$ the SIR algorithm draws N independent samples from $\hat\pi_{n-1}^\mu$, namely,

$$Z_{n-1}(i) \sim \hat\pi_{n-1}^\mu \qquad i \in \{1,\dots,N\},$$

and then samples

$$X_n(i) \sim P(Z_{n-1}(i), \cdot\,) \qquad i \in \{1,\dots,N\}.$$

Then

$$\mathsf{S}^N\mathsf{P}\hat\pi_{n-1}^\mu = \frac{1}{N}\sum_{i=1}^N \delta_{X_n(i)},$$

and by applying C_n we finally get

$$\hat\pi_n^\mu := \mathsf{C}_n\mathsf{S}^N\mathsf{P}\hat\pi_{n-1}^\mu = \sum_{i=1}^N W_n(i)\,\delta_{X_n(i)},$$

where

$$W_n(i) := \frac{g(X_n(i), Y_n)}{\sum_{\ell=1}^N g(X_n(\ell), Y_n)} \qquad i \in \{1,\dots,N\}. \tag{3.4}$$

Instead of repeatedly updating the weights as in the SIS algorithm, cf. (3.2), the SIR algorithm resets all the weights to $1/N$ at each iteration, before updating them in the correction step using the likelihood of the new observation. The implementation of the algorithm is described in Figure 3.4.

The process of sampling from the distribution $\hat\pi_{n-1}^\mu$ is usually referred to as the *resampling step*, as N particles are sampled from an empirical measure that is itself defined via N particles, specifically,

$$\hat\pi_{n-1}^\mu = \sum_{i=1}^N W_{n-1}(i)\,\delta_{X_{n-1}(i)} \qquad X_{n-1}(1),\dots,X_{n-1}(N) \text{ are i.i.d.} \sim \mathsf{P}\hat\pi_{n-2}^\mu.$$

[1] In the SIR algorithm the sampling operator is applied iteratively in time. At each iteration of the algorithm, samples are drawn conditionally independent given the collection of all random variables generated by the algorithm up to that iteration.

Algorithm 2: SIR particle filter / Bootstrap particle filter

Data: Fix $n, N \geq 1$. Let the observations $Y_1, \ldots, Y_n$ be given.
Let $\hat{\pi}_0^\mu = \mu$;
for $k = 1, \ldots, n$ **do**
  Sample i.i.d. $Z_{k-1}(i)$, $i = 1, \ldots, N$ from the distribution $\hat{\pi}_{k-1}^\mu$;
  Sample $X_k(i) \sim p(Z_{k-1}(i), \cdot)\, d\psi$, $i = 1, \ldots, N$;
  Compute $W_k(i) = g(X_k(i), Y_k) / \sum_{\ell=1}^N g(X_k(\ell), Y_k)$, $i = 1, \ldots, N$;
  Let $\hat{\pi}_k^\mu = \sum_{i=1}^N W_k(i)\, \delta_{X_k(i)}$;
Compute the approximate filter $\pi_n^\mu f \approx \hat{\pi}_n^\mu f$.

Figure 3.4: The classical sequential importance resampling (SIR) particle filter.

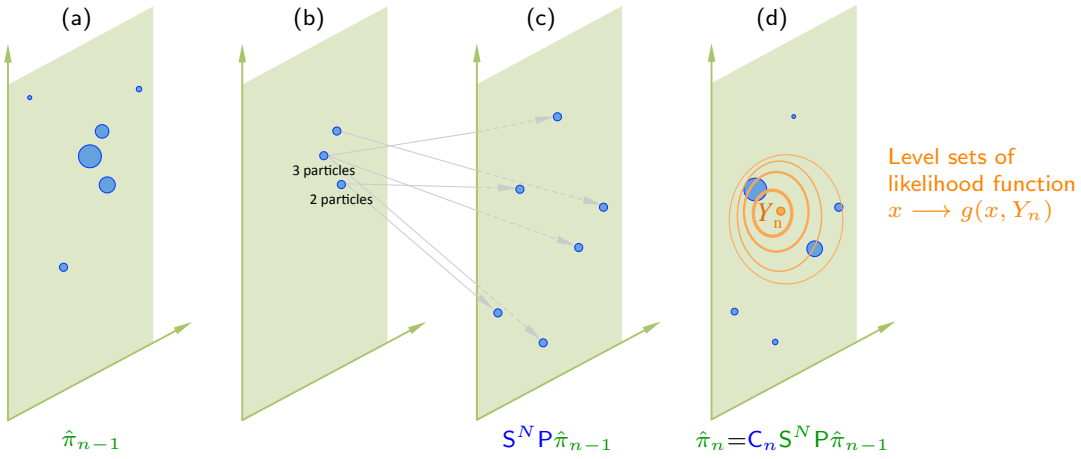

Figure 3.5: Representation of a single iteration of the SIR particle filter with $\mathbb{X} = \mathbb{Y} = \mathbb{R}_+^2$ and $N = 6$. Each particle is represented by a blue ball, whose size is proportional to the weight of the particle. (a) Representation of $\hat{\pi}_{n-1}$. (b) Resampling step: N particles are sampled independently with replacement and weights are reset to $1/N$. If a number m is attached to a particle, then there are m particles sharing the same location. (c) Particles are propagated forward using the underlying dynamics. (d) Particles are reweighed according to the likelihood of the new observation at time n (whose level sets are drawn in orange) yielding $\hat{\pi}_n$, following the weight recursion (3.4).

In the resampling step particles with with low weights are less likely to be sampled than particles with high weights. So, in the resampling step some of the particles with low weights will disappear, while particles with large weights will be sampled more than once. Figure 3.5 illustrates a typical iteration of the algorithm.

The resampling step is the basic mechanism that allows the SIR algorithm to overcome the weight impoverishment problem of the SIS algorithm with time (Section 3.2.1). In the next section we make this intuition precise by providing a detailed error analysis for the SIR particle filter.

3.3.1 Filter stability and time-uniform error bounds

While the convergence analysis for the SIS particle filter is straightforward as the algorithm is defined in terms of a collection of independent particles, for the SIR algorithm the situation is more involved as at each iteration the resampling step introduces dependency among particles (for example, recall that particles with high weights are likely to be duplicated). Nonetheless, it is easily shown that for each $n \geq 1$ the particle filter $\hat{\pi}_n^\mu$ converges to the exact filter π_n^μ as $N \to \infty$. To gain some insight into the approximation properties of the SIR particle filter, let us perform the simplest possible error analysis. Recall from Section 2.3 the following distance between (possibly random) probability measures ρ, ρ' on $\mathbb{X}$:

$$|||\rho - \rho'||| := \sup_{|f| \leq 1} \sqrt{\mathbf{E}\,(\rho f - \rho' f)^2}.$$

From Lemma 2.10 and Lemma 2.17 we have

$$|||\mathsf{P}\rho - \mathsf{P}\rho'||| \leq |||\rho - \rho'|||, \qquad |||\rho - \mathsf{S}^N \rho||| \leq \frac{1}{\sqrt{N}}.$$

Let us assume for simplicity that the observation density g is bounded away from zero and infinity, that is, $\kappa \leq g(x, y) \leq \kappa^{-1}$ for some $0 < \kappa < 1$. From Lemma 2.9 (choosing $g(x) := g(x, Y_n)$) we obtain

$$|||\mathsf{C}_n\rho - \mathsf{C}_n\rho'||| \leq 2\kappa^{-2}|||\rho - \rho'|||.$$

Putting these bounds together and using the triangle inequality for the metric $|||\cdot|||$ we find

$$\begin{aligned} |||\mathsf{F}_n\rho - \hat{\mathsf{F}}_n\rho'||| = |||\mathsf{C}_n\mathsf{P}\rho - \mathsf{C}_n\mathsf{S}^N\mathsf{P}\rho'||| &\leq 2\kappa^{-2}\left\{|||\mathsf{P}\rho - \mathsf{P}\rho'||| - |||\mathsf{P}\rho' - \mathsf{S}^N\mathsf{P}\rho'|||\right\} \\ &\leq 2\kappa^{-2}\left\{|||\rho - \rho'||| + \frac{1}{\sqrt{N}}\right\}. \end{aligned}$$

By iterating this inequality n times, using that $\pi_0^\mu = \hat{\pi}_0^\mu$, we find

$$|||\pi_n^\mu - \hat{\pi}_n^\mu||| \leq 2\kappa^{-2}\left\{|||\pi_{n-1}^\mu - \hat{\pi}_{n-1}^\mu||| + \frac{1}{\sqrt{N}}\right\} \leq \frac{C_n}{\sqrt{N}},$$

with

$$C_n := \sum_{i=1}^{n} (2\kappa^{-2})^i.$$

So, for a fixed time $n \geq 1$ the bootstrap particle filter does indeed approximate the exact nonlinear filter as the number of particles N goes to infinity, with the typical Monte Carlo $1/\sqrt{N}$-rate.

In many applications, however, one needs to have good estimates for the filter at arbitrary times. This is the case, for instance, of target tracking, where the goal is to continuously track the location of the target. The analysis that we have performed so far does not guarantee that the SIR particle filter can be successfully applied to this

end, as C_n grows exponentially in time n. Fortunately, the exponential growth of the error is an artifact of our crude bound and typically does not occur in practice. The reason why the constant C_n obtained above growths with time is that we have performed a *recursive* error analysis of the algorithm: we bounded the error committed at each time step, and we naively iterated this bound for n steps, so that the error accumulates over time.

We presently show that a more refined analysis that exploits the behavior of the filter distribution itself—instead of working at the level of the filter recursion—yields the following time-uniform error bound:

$$\sup_{n\geq 0} |||\pi_n^\mu - \hat{\pi}_n^\mu||| \leq \frac{C}{\sqrt{N}},$$

where C is a constant that does not depend on time. This is the reason why the SIR algorithm has proved to perform extraordinarily well in many classical applications such as target tracking, speech recognition, and finance [8].

The property of the filter that allows this analysis is the so-called *filter stability* property, which roughly says that π_n^μ forgets its initial condition μ as $n \to \infty$. As first realized by Del Moral and Guionnet in 2001 [15], the stability property provides a dissipation mechanism that mitigates the accumulation of approximation errors over time, yielding time-uniform error bounds. In the reminder of this section we make this idea precise under certain (strong) conditions.

Recall that both the filter and the SIR particle filter are defined recursively:

$$\pi_n^\mu := \mathsf{F}_n \cdots \mathsf{F}_1 \mu, \qquad \hat{\pi}_n^\mu := \hat{\mathsf{F}}_n \cdots \hat{\mathsf{F}}_1 \mu, \qquad n \geq 1,$$

where $\mathsf{F}_n := \mathsf{C}_n\mathsf{P}$, $\hat{\mathsf{F}}_n := \mathsf{C}_n\mathsf{S}^N\mathsf{P}$, and $\pi_0^\mu = \hat{\pi}_0^\mu = \mu$. The basic idea that allows to prove time-uniform bounds for the bootstrap particle filter is based on the following simple error decomposition [8]. If we write $\pi_n^\mu - \hat{\pi}_n^\mu$ as a telescoping sum:

$$\pi_n^\mu - \hat{\pi}_n^\mu = \sum_{s=1}^n \{\mathsf{F}_n \cdots \mathsf{F}_{s+1}\mathsf{F}_s\hat{\mathsf{F}}_{s-1}\cdots\hat{\mathsf{F}}_1\mu - \mathsf{F}_n \cdots \mathsf{F}_{s+1}\hat{\mathsf{F}}_s\hat{\mathsf{F}}_{s-1}\cdots\hat{\mathsf{F}}_1\mu\},$$

then by the triangle inequality we get

$$|||\pi_n^\mu - \hat{\pi}_n^\mu||| \leq \sum_{s=1}^n |||\mathsf{F}_n \cdots \mathsf{F}_{s+1}\mathsf{F}_s\hat{\pi}_{s-1}^\mu - \mathsf{F}_n \cdots \mathsf{F}_{s+1}\hat{\mathsf{F}}_s\hat{\pi}_{s-1}^\mu|||. \tag{3.5}$$

The s-th term in this sum could be interpreted as the contribution to the total error at time n due to the filter approximation made at time s. The key insight is now that one can employ the filter stability property to control this sum uniformly in time.

The following theorem establishes filter stability in its simplest form, under a certain ergodicity assumption on the signal process called *mixing condition*. As shown in Lemma 2.10, this condition causes the signal $(X_n)_{n\geq 0}$ itself to forget its initial condition at an exponential rate, and the following results shows how the filter inherits this property.

Theorem 3.7 (Filter stability, inheritance)**.** *Suppose that the transition density p satisfies the following* mixing condition*: there exists a constant $0 < \varepsilon < 1$ such that*

$$\varepsilon \le p(x,z) \le \varepsilon^{-1} \qquad \text{for all } x, z \in \mathbb{X}.$$

Then, for any two (possibly random) probability measures ρ and ρ' on $(\mathbb{X}, \mathcal{X})$ we have, for $n > s$,

$$|||\mathsf{F}_n \cdots \mathsf{F}_{s+1}\rho - \mathsf{F}_n \cdots \mathsf{F}_{s+1}\rho'||| \le 2\,\varepsilon^{-2}(1-\varepsilon^2)^{n-s}|||\rho - \rho'|||.$$

Proof. For each $1 \le k \le n$ define the (random) transition kernel

$$K_{k|n}(x, A) := \mathbf{P}(X_k \in A | X_{k-1} = x, Y_{1:n}).$$

Proceeding as in the proof of Lemma 3.3, as $K_{k|n}(x, \cdot)$ is the marginal of the distribution $\mathbf{P}(X_{k:n} \in \cdot\,|X_{k-1} = x, Y_{1:n})$ on the X_k coordinate, it is easy to verify that

$$K_{k|n}(x, A) = \frac{\int p(x,x')\,\psi(dx')\,\beta_{k|n}(x', Y_{k+1:n})\,g(x', Y_k)\,\mathbf{1}_A(x')}{\int p(x,x')\,\psi(dx')\,\beta_{k|n}(x', Y_{k+1:n})\,g(x', Y_k)},$$

where $\beta_{k|n}$ can be defined through the backward recursion

$$\beta_{k|n}(x, Y_{k+1:n}) := \int p(x,x')\psi(dx')\,g(x', Y_{k+1})\,\beta_{k+1|n}(x', Y_{k+2:n}), \qquad \beta_{n|n} := 1.$$

By the Markov property it is easy to verify that conditionally on $Y_1, \ldots, Y_n$ the random variables $X_0, \ldots, X_n$ follows the law of a Markov chain. In fact, for each $1 \le k \le n$ we have

$$\mathbf{P}(X_k \in A | X_{0:k-1}, Y_{1:n}) = K_{k|n}(X_{k-1}, A)$$

and for any probability measure ρ on $(\mathbb{X}, \mathcal{X})$ and any real-valued measurable function f we have

$$\begin{aligned}
(\mathsf{F}_n \cdots \mathsf{F}_1\rho)f &= \int \mathbf{P}^\rho(X_{0:n} \in dx_{0:n} | Y_{1:n})\, f(x_n) \\
&= \int \mathbf{P}^\rho(X_0 \in dx_0 | Y_{1:n}) \prod_{k=1}^{n} \mathbf{P}^\rho(X_k \in dx_k | X_{0:k-1} = x_{0:k-1}, Y_{1:n})\, f(x_n) \\
&= \rho_{0|n} K_{1|n} \cdots K_{n|n} f,
\end{aligned}$$

where we have defined $\rho_{0|n} := \mathbf{P}^\rho(X_0 \in \cdot\,|Y_{1:n})$. By the same argument, as $\mathsf{F}_n \cdots \mathsf{F}_1$ and $\mathsf{F}_n \cdots \mathsf{F}_{s+1}$, for any $0 \le s < n$, differ only in that a different sequence of observations ($Y_1, \ldots, Y_n$ versus $Y_{s+1}, \ldots, Y_n$) is used in the computation of these quantities, we have

$$\mathsf{F}_n \cdots \mathsf{F}_{s+1}\rho = \rho_{s|n} K_{s+1|n} \cdots K_{n|n},$$

and it is easy to check that

$$\rho_{s|n}(A) := \frac{\int \rho(dx)\,\beta_{s|n}(x, Y_{s+1:n})\,\mathbf{1}_A(x)}{\int \rho(dx)\,\beta_{s|n}(x, Y_{s+1:n})}.$$

Therefore, by Lemma 2.10 and Lemma 2.9 we have

$$\begin{aligned}|||\mathsf{F}_n \cdots \mathsf{F}_{s+1}\rho - \mathsf{F}_n \cdots \mathsf{F}_{s+1}\rho'||| &= |||\rho_{s|n}K_{s+1|n}\cdots K_{n|n} - \rho'_{s|n}K_{s+1|n}\cdots K_{n|n}||| \\ &\leq (1-\varepsilon^2)^{n-s}|||\rho_{s|n} - \rho'_{s|n}||| \\ &\leq 2\,\frac{\sup_{x\in\mathbb{X}}\beta_{s|n}(x, Y_{s+1:n})}{\inf_{x\in\mathbb{X}}\beta_{s|n}(x, Y_{s+1:n})}\,(1-\varepsilon^2)^{n-s}|||\rho - \rho'|||.\end{aligned}$$

The proof is immediately concluded once we notice that by the mixing conditions we have

$$\varepsilon C \leq \beta_{s|n}(x, Y_{s+1:n}) \leq \varepsilon^{-1}C,$$

where

$$C := \int g(x', Y_{s+1})\,\beta_{s+1|n}(x', Y_{s+2:n}).$$

□

Under the mixing condition for the signal, Theorem 3.7 tells us that the filter forgets its initial condition at a geometric rate. This also means that past approximation errors are forgotten at an exponential rate: if we substitute the stability property in the error decomposition (3.5), we obtain

$$|||\pi_n^\mu - \hat{\pi}_n^\mu||| \leq \sum_{s=1}^{n} 2\,\varepsilon^{-2}(1-\varepsilon^2)^{n-s}|||\mathsf{F}_s\hat{\pi}_{s-1}^\mu - \hat{\mathsf{F}}_s\hat{\pi}_{s-1}^\mu||| \leq 2\,\varepsilon^{-4}\sup_{n,\rho}|||\mathsf{F}_n\rho - \hat{\mathsf{F}}_n\rho|||.$$

Thus, if we can control the error $|||\mathsf{F}_n\rho - \hat{\mathsf{F}}_n\rho|||$ in a single time step, we obtain a time-uniform bound of the same order. In the case of the bootstrap particle filter, if $\kappa \leq g(x,y) \leq \kappa^{-1}$, we have that

$$|||\mathsf{F}_n\rho - \hat{\mathsf{F}}_n\rho||| = |||\mathsf{C}_n\mathsf{P}\rho - \mathsf{C}_n\mathsf{S}^N\mathsf{P}\rho||| \leq \frac{2\kappa^{-2}}{\sqrt{N}},$$

and we obtain a time-uniform version of the crude error bound:

$$\sup_{n\geq 0}|||\pi_n^\mu - \hat{\pi}_n^\mu||| \leq 4\,\varepsilon^{-4}\kappa^{-2}\,\frac{1}{\sqrt{N}}.$$

Let us remark at this point that the basic error decomposition discussed above allows us to separate the problem of obtaining time-uniform bounds into two parts: the one-step approximation error and the stability property. The development of these ingredients constitutes the bulk of the framework that is introduced in Chapter 4 to deal with filtering problems in high dimension.

Remark 3.8 (Results in the literature)**.** *In [15] Del Moral and Guionnet prove several time-uniform error bounds for the SIR algorithm, under assumptions on filter stability that are also weaker compared the one considered in Theorem 3.7. Presently, we limit our treatment the basic ideas that are instrumental for the framework that will be developed in Chapter 4.*

3.3.2 The curse of dimensionality.

While the SIR algorithm provides estimates that have error bounds uniform with time, it turns out that this algorithm suffers severely from the curse of dimensionality with respect to the spatial dimension of the model. It is far from obvious at this point why this should be the case. Indeed, the state spaces $\mathbb{X}$ and $\mathbb{Y}$ have only been assumed to be Polish (a mild technical assumption meant only to ensure the existence of regular conditional probabilities), and no explicit notion of dimension appears in the above error bound. To understand why the bound

$$\sup_{n\geq 0} |||\pi_n^\mu - \hat{\pi}_n^\mu||| \leq \frac{C}{\sqrt{N}}$$

is typically exponential in the model dimension, we must consider a suitable class of high-dimensional models in which the dependence on dimension can be explicitly investigated. In the present section we consider a simple class of *trivial* high-dimensional models that is useless in any application, but is nonetheless helpful for the purpose of developing intuition for dimensionality issues in particle filters. Moreover, this trivial class of models represents the backbone of the more realistic framework that will be considered in the next two chapters (see Section 4.1).

In a d-dimensional model, X_n and Y_n are each described by d coordinates: X_n^i, Y_n^i, $i \in \{1, \ldots, d\}$. To construct a trivial d-dimensional model, we simply start with a given one-dimensional model and duplicate it d times. That is, let $(\tilde{X}_n, \tilde{Y}_n)_{n\geq 0}$ be a hidden Markov model on $\tilde{\mathbb{X}} \times \tilde{\mathbb{Y}}$ with transition density $\tilde{p}$ and observation density $\tilde{g}$ with respect to reference measures $\tilde{\psi}$ and $\tilde{\varphi}$, respectively. Then we set

$$\mathbb{X} = \tilde{\mathbb{X}}^d, \qquad \mathbb{Y} = \tilde{\mathbb{Y}}^d, \qquad \psi = \tilde{\psi}^{\otimes d}, \qquad \varphi = \tilde{\varphi}^{\otimes d},$$

and

$$p(x, z) = \prod_{i=1}^{d} \tilde{p}(x^i, z^i), \qquad g(x, y) = \prod_{i=1}^{d} \tilde{g}(x^i, y^i),$$

so that each coordinate $(X_n^i, Y_n^i)_{n\geq 0}$ is an independent copy of $(\tilde{X}_n, \tilde{Y}_n)_{n\geq 0}$. The (trivial) dependency structure of this model is represented in Figure 3.6. Note that we have used the term d-dimensional in the sense that our model has d independent degrees of freedom: each degree of freedom can itself in principle take values in a high- or even infinite-dimensional state space $\tilde{\mathbb{X}} \times \tilde{\mathbb{Y}}$. This is, however, precisely the notion of dimension that is relevant to the curse of dimensionality (in [4, 47] this idea is sharpened by a notion of "effective dimension").

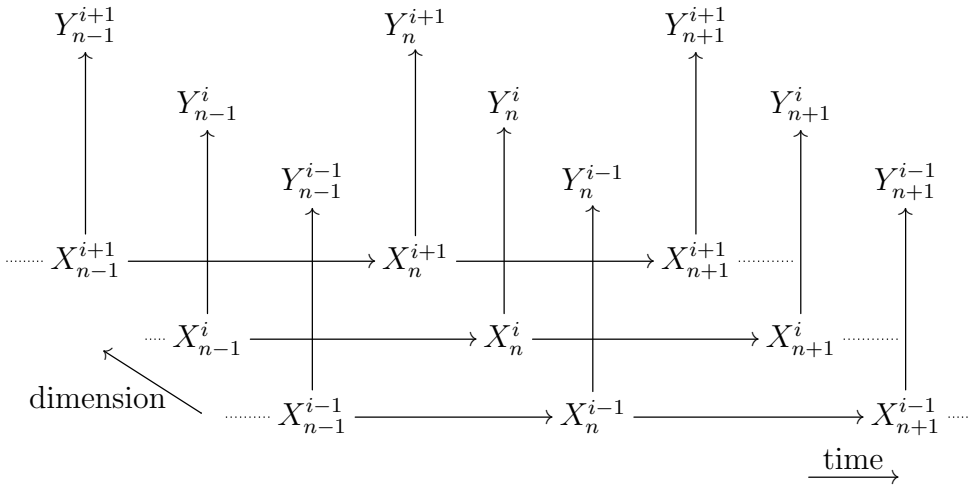

Figure 3.6: Dependency graph of a (trivial) high-dimensional filtering model.

In this trivial setting, it is now easily seen how the curse of dimensionality arises in our error bound. Indeed, let us assume again for simplicity that $\kappa \le \tilde{g}(\tilde{x},\tilde{y}) \le \kappa^{-1}$ for some $0 < \kappa < 1$. Then $\kappa^d \le g(x,y) \le \kappa^{-d}$, so we obtain a bound that is exponential in the dimension d even after only one time step:

$$|||\pi_1^\mu - \hat{\pi}_1^\mu||| \le \frac{2\kappa^{-2d}}{\sqrt{N}}.$$

An inspection of our bound clarifies the source of this exponential growth: even though the Monte Carlo sampling itself is dimension-free ($|||\rho - \mathsf{S}^N\rho||| \le N^{-1/2}$ independent of dimension, see Lemma 2.17), the correction operator C_n, which is highly nonlinear, blows up the sampling error exponentially in high dimension. In particular, it is evidently the dimension of the observations, rather than that of the underlying model, that controls the exponential growth in our error bound.

Of course, the above analysis is far from convincing. First of all, we have only proved a rather crude upper bound on the approximation error, so that it might be possible that a more sophisticated bound would eliminate the exponential dependence on dimension as was done using the filter stability property to eliminate the exponential dependence on time. Second, one could argue that our strong notion of approximation with respect to the $|||\cdot|||$-norm is too restrictive to give meaningful results in high dimension (which is in fact the case: we will later consider *local* error bounds instead), so that a weaker notion of approximation might avoid the exponential dependence on dimension. Unfortunately, the much more delicate analysis of Bickel *et al.* [4, 47] demonstrates conclusively that the curse of dimensionality of the bootstrap particle filter is a genuine phenomenon and not a mathematical deficiency of our analysis, as we will briefly explain presently. Nonetheless, both the ideas raised above to eliminate the exponential dependence on dimension will play an important role in the framework developed in Chapter 4.

3.3.3 Sample degeneracy with dimension

The reason why the SIR algorithm performs poorly when the model dimension is high is essentially the same reason why the SIS algorithm behaves badly when the time-horizon is large, and it has to do with the fact that the importance sampling paradigm is typically very inefficient in high-dimensional models. As the SIS algorithm approximates the smoothing distribution $\mathbf{P}^\mu(X_{0:n} \in \cdot\,|Y_{1:n})$, the dimension of interest in that case is time: weight degeneracy occurs as n increases[2]. On the other hand, in the current analysis of the SIR algorithm in the trivial model at hand, the dimension of interest is the number of hidden Markov chains in the model: weight degeneracy occurs as d increases, and it is manifested even in a single iteration of the algorithm, as the following two examples illustrate.

This example represents the analog of Example 3.4 for the SIR algorithm.

Example 3.9 (Weight degeneracy of SIR with dimension)**.** *In the framework introduced in Section 3.3.2, consider the hidden Markov model where $(X_n)_{n\geq 0}$ is a symmetric random walk in $\mathbb{Z}^d$, $d \geq 1$, with $X_0 = x \in \mathbb{Z}^d$, and for each $n \geq 1$ we have $Y_n = X_n + \varepsilon\eta_n$, where $\varepsilon \in \mathbb{R}_+$ and $(\eta_n)_{n\geq 0}$ is a collection of i.i.d. random variables having the standard Gaussian distribution in $\mathbb{R}^d$ (that is, zero mean and identity covariance matrix). We now look at the first iteration of the SIR algorithm. If the signal-to-noise ratio is high, that is, if ε is very close to 0, then we expect the distribution $\mathbf{P}^x(X_1 \in \cdot\,|Y_1 = y_1)$ to be very concentrated around X_1, the true location of the signal at time 1. However, if we sample N particles from the distribution $\mathbf{P}^x(X_1 \in \cdot\,)$, then on average only $N/2^d$ particles will be close to X_1, and the weight degeneracy gets exponentially worse as the dimension d increases. Figure 3.7 represents this scenario.*

The following asymptotical analysis (in the limit $N \to \infty$) gives another quick illustration of the degeneracy in dimension of the SIR algorithm. This example is the analog of Example 3.5 in space.

Example 3.10 (Exponential growth of the SIR asymptotic variance with dimension)**.** *Consider the (trivial) d-dimensional model introduced in Section 3.3.2. Let $\tilde{\mu}$ be a probability measure on $\tilde{\mathbb{X}}$, and define $\mu = \tilde{\mu}^{\otimes d}$ on $\mathbb{X}$. Let f be a measurable function on $\mathbb{X}$ such that $f(x) = f(\tilde{x})$ whenever $x^\ell = \tilde{x}^\ell$, for a certain $\ell \in \{1, \dots, d\}$. Then,*

$$N^{1/2}(\hat{\pi}_1^\mu f - \pi_1^\mu f) \xrightarrow{\text{in distribution}} \text{Gaussian}\left(0, \sigma_d^2(f)\right) \qquad \text{as } N \to \infty,$$

with

$$\sigma_d^2(f) := c(f)\,\gamma^{d-1},$$

where $c(f)$ and γ are constants that do not depend on d, $c(f) > 0$ as long as f is not a constant, and $\gamma > 1$ as long as the observation density $\tilde{g}$ is different from 1.

[2]Note that in our analysis of the SIS algorithm we ignored the curse of dimensionality with respect to the model dimension. This issue is exactly the same as for the SIR algorithm, as this type of weight degeneracy already appears in one iteration of the SIR particle filter. See Section 3.2.1.

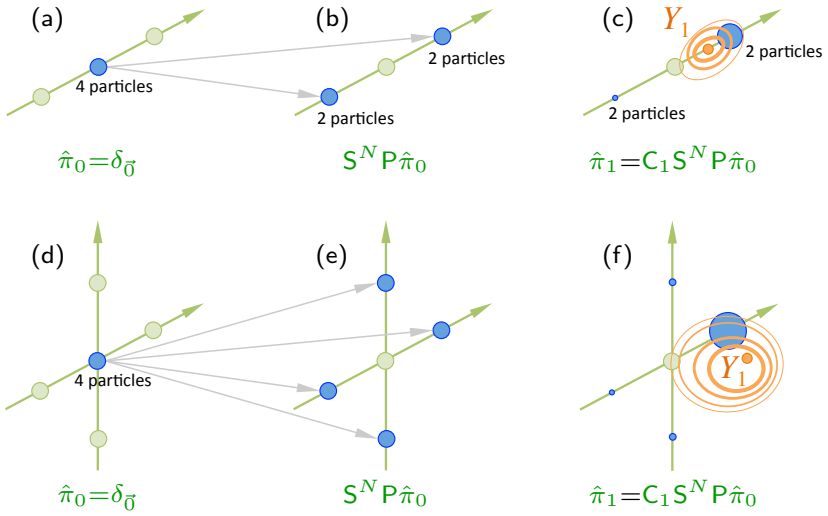

Figure 3.7: Representation of the first iteration of the SIR particle filter applied to the hidden Markov model described in Example 3.9, with $N = 4$ particles and $x = \vec{0}$. Pictures (a), (b) and (c) refer for the case $d = 1$, while pictures (d), (e) and (f) refer for the case $d = 2$. Each particle is represented by a blue ball, whose size is proportional to the weight of the particle, and orange curves represents the level sets of thelikelihood function. As symbolically represented, after the first iteration of the algorithm only an average of $N/2^d$ particles have meaningful weights, which is a manifestation of the curse of dimensionality.

For each $x \in \tilde{\mathbb{X}}, y \in \tilde{\mathbb{Y}}$ define

$$\bar{g}(x, y) := \frac{\tilde{g}(x, y)}{\int \tilde{\mu}(dz)\, \tilde{p}(z, x)\, \tilde{\psi}(dx)\, \tilde{g}(x, y)}.$$

Then, for each $N \geq 1$ we have

$$N^{1/2}(\hat{\pi}_1^\mu f - \pi_1^\mu f) = \frac{N^{-1/2} \sum_{i=1}^N (f(X_1(i)) - \pi_1^\mu f) \prod_{k=1}^d \bar{g}(X_1^k(i), Y_1^k)}{N^{-1} \sum_{i=1}^N \prod_{k=1}^d \bar{g}(X_1^k(i), Y_1^k)}, \tag{3.6}$$

where $(X_1(i))_{i=1,\dots,N}$ is a collection of i.i.d. random variables with distribution $(\mathsf{P}\mu)(A) = \int \prod_{k=1}^d \tilde{\mu}(dz^k) \tilde{p}(z^k, x^k) \tilde{\psi}(dx^k)\, \mathbf{1}_A(x)$, conditionally independent given Y_1. By independence, for each i we have

$$\begin{aligned} \mathbf{E} \prod_{k=1}^d \bar{g}(X_1^k(i), Y_1^k) &= \prod_{k=1}^d \mathbf{E}\, \mathbf{E}(\bar{g}(X_1^k(i), Y_1^k) | Y_1^k) \\ &= \prod_{k=1}^d \mathbf{E} \int \tilde{\mu}(dz)\, \tilde{p}(z, x)\, \tilde{\psi}(dx)\, \bar{g}(x, Y_1^k) = 1, \end{aligned}$$

and the strong Law of Large Numbers yields, as $N \to \infty$,

$$N^{-1} \sum_{i=1}^N \prod_{k=1}^d \bar{g}(X_1^k(i), Y_1^k) \xrightarrow{\text{almost surely}} \mathbf{E} \prod_{k=1}^d \bar{g}(X_1^k(i), Y_1^k) = 1$$

for the denominator in (3.6). *On the other hand, as*

$$\pi_1^\mu f = \int \prod_{k=1}^{d} \tilde{\mu}(dz^k)\, \tilde{p}(z^k, x^k)\, \tilde{\psi}(dx^k)\, \bar{g}(x^k, Y_1^k)\, f(x),$$

by independence it is immediately verified that

$$\mathbf{E}\left(f(X_1(i)) - \pi_1^\mu f\right) \prod_{k=1}^{d} \bar{g}(X_1^k(i), Y_1^k) = 0$$

and

$$\sigma_d^2(f) := \mathbf{E}\left(\left(f(X_1(i)) - \pi_1^\mu f\right) \prod_{k=1}^{d} \bar{g}(X_1^k(i), Y_1^k)\right)^2 = c(f)\, \gamma^{d-1},$$

where

$$c(f) := \mathbf{E} \int \tilde{\mu}(dz^\ell)\, \tilde{p}(z^\ell, x^\ell)\, \tilde{\psi}(dx^\ell)\, \left(f(x) - \pi_1^\mu f\right)^2 \bar{g}(x^\ell, Y_1^\ell)^2,$$

$$\gamma := \mathbf{E} \int \tilde{\mu}(dz)\, \tilde{p}(z, x)\, \tilde{\psi}(dx)\, \bar{g}(x, Y_1^1)^2.$$

The Central Limit Theorem yields that the numerator in (3.6) *converges in distribution as* $N \to \infty$ *to a Gaussian distribution with mean* 0 *and variance* $\sigma_d^2(f)$. *Therefore, it is immediate that also* (3.6) *converges in distribution to the same Gaussian distribution. Applying Jensen's inequality twice we immediately get that* $\gamma > 1$ *as long as* $\tilde{g}$ *is different from* 1.

The key obstacle when the observations are high-dimensional is that the posterior measure $\mathsf{C}_n\rho$ is nearly singular with respect to the prior measure ρ (cf. Proposition 2.14). In particular, a point that has high likelihood under ρ has likelihood under $\mathsf{C}_n\rho$ that is exponentially small in the dimension. Therefore, if we draw a fixed number N of samples from ρ, then with very high probability every one of these samples will have exponentially small likelihood under $\mathsf{C}_n\rho$ and, as is common in rare-event scenarios, the least unlikely sample will be exponentially more likely than any of the other samples. Thus $\mathsf{C}_n\mathsf{S}^N\rho$ will put almost all its mass on the sample with the largest likelihood, which yields effectively a Monte Carlo approximation of $\mathsf{C}_n\rho$ with sample size 1 rather than N. This situation is illustrated in Figure 3.8. This weight degeneracy phenomenon rules out any meaningful form of approximation in high dimension. In [4, 47], a careful analysis shows that the collapse phenomenon occurs unless the sample size N is taken to be exponential in the dimension, which provides a rigorous statement of the curse of dimensionality.

Remark 3.11. *(Curse of dimensionality and sample degeneracy) Sample degeneracy is the manifestation of the curse of dimensionality phenomenon in particle filters, but it does not coincides with it. For instance, particle degeneracy appears also in low dimensional models if the noise driving both the dynamics and the observation is low [61].*

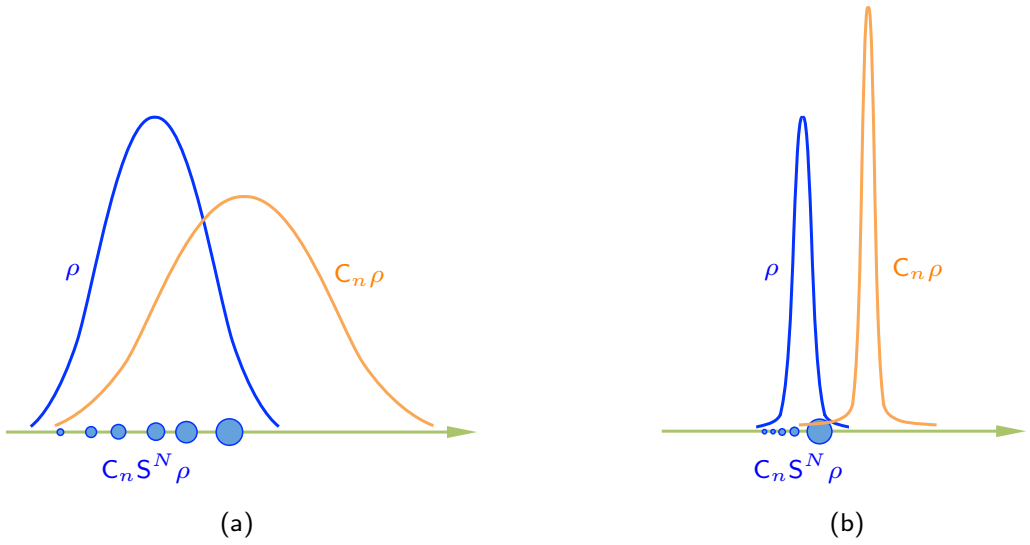

Figure 3.8: Illustration of weights degeneracy with model dimension in a typical iteration of the SIR particle filter. (a) Probability measures in low dimension. (b) Probability measures in high dimension (low-dimensional representation). In high dimension ρ and $\mathsf{C}_n\rho$ tend to put mass on different portions of the space. This is the reason why already after a single iteration of the SIR particle filter only a small fraction of samples (in fact, a fraction that is exponentially small in the dimension) is relevant in the algorithm. Each sample X from ρ is represented by a blue ball whose size is proportional to the likelihood $g(X, Y_n)$, as prescribed by the weights definition (3.4).

Despite that the SIR particle filter suffers from the curse of dimensionality when applied to the full (trivial) model of Section 3.3.2, it is obvious in this case that one can surmount this problem in a simple fashion: as each of the coordinates of the high-dimensional model is independent, one can simply run an independent SIR filter in each coordinate. It is evident that the local error of this algorithm (that is, the error of the marginal of the filter in each coordinate) is, by construction, independent of the model dimension d. In this sense, this trivial model shows that it is indeed possible to filter very efficiently regardless of the ambient dimension (though not with the SIR particle filter, which fails spectacularly). Chapter 4 builds on this intuition by considering a more general class of models and by developing a sampling strategy that can overcome the weights degeneracy with model dimension.

Remark 3.12 (Smoothing in high dimension)**.** *If, instead of computing the filter* $\mathbf{P}(X_n \in \cdot\,|Y_1, \ldots, Y_n)$*, we wish to compute the full conditional path distribution* $\mathbf{P}(X_0, \ldots, X_n \in \cdot\,|Y_1, \ldots, Y_n)$ *(known as the smoothing problem), then Markov Chain Monte Carlo (MCMC) methods can be successfully employed in high dimension. However, this procedure requires the entire history of observations and is not recursive, so that it cannot be implemented on-line and is impractical over a long time horizon (cf. [3]). The crucial question to be addressed is therefore whether it is possible to develop filtering algorithms that are both recursive and that admit error bounds that are uniform in time and in the model dimension.*

Remark 3.13 (Importance sampling)**.** *As in the case of the SIS algorithm (cf. Remark 3.6), also the SIR algorithm can be described as an instance of the self-normalized importance sampling paradigm introduced in Section 2.5, and different importance distributions can be considered. While the practical performance of the SIR algorithm can be largely improved by working with importance distributions that are tailored to the specific model being investigated, the benefit is limited to reducing the constants sitting in front of the error bounds, and this technique does not provide a fundamental solution to the curse of dimensionality. A new paradigm is needed, as we will see in the next chapter.*

Presently we link our formulation of the SIR algorithm with the one usually considered in the literature (see [19] for instance). First of all, notice the following identity[3] *which holds for each $n, N \geq 1$, and for each probability measure ρ on $(\mathbb{X}, \mathcal{X})$:*

$$\mathsf{C}_n \mathsf{S}^N \rho = \mathsf{S}^N_\rho \mathsf{C}_n \rho.$$

In fact, by definition of C_n and S^N we have

$$\mathsf{C}_n \mathsf{S}^N \rho = \frac{\sum_{i=1}^N g(X(i), Y_n)\, \delta_{X(i)}}{\sum_{i=1}^N g(X(i), Y_n)}, \qquad X(1), \dots, X(N) \text{ are i.i.d. samples} \sim \rho.$$

On the other hand, as the Radon-Nikodym derivative between $\mathsf{C}_n\rho$ and ρ reads

$$\frac{d(\mathsf{C}_n \rho)}{d\rho}(x) = \frac{g(x, Y_n)}{\int \rho(dx)\, g(x, Y_n)},$$

from the definition of S^N_ρ (Definition 2.18) we have

$$\mathsf{S}^N_\rho \mathsf{C}_n \rho = \frac{\sum_{i=1}^N \frac{d(\mathsf{C}_n\rho)}{d\rho}(X(i))\, \delta_{X(i)}}{\sum_{i=1}^N \frac{d(\mathsf{C}_n\rho)}{d\rho}(X(i))} = \mathsf{C}_n \mathsf{S}^N \rho.$$

Therefore, the SIR algorithm introduced in the main text can formulated as follows:

$$\hat\pi^\mu_{n-1} \xrightarrow{\text{prediction}} \mathsf{P}\hat\pi^\mu_{n-1} \xrightarrow{\text{correction}} \mathsf{C}_n\mathsf{P}\hat\pi^\mu_{n-1} \xrightarrow{\substack{\text{importance}\\ \text{sampling}}} \hat\pi^\mu_n := \mathsf{S}^N_{\lambda_n}\mathsf{C}_n\mathsf{P}\hat\pi^\mu_{n-1},$$

where the importance distribution is $\lambda_n = \mathsf{P}\hat\pi^\mu_{n-1}$.

The so-called "optimal" distribution is given by the choice $\lambda^\star_n = \mathsf{C}_n\mathsf{P}\hat\pi^\mu_{n-1}$. As $\mathsf{S}^N_\mu \mu = \mathsf{S}^N \mu$, this choice yields the following algorithm

$$\hat\pi^\mu_{n-1} \xrightarrow{\text{prediction}} \mathsf{P}\hat\pi^\mu_{n-1} \xrightarrow{\text{correction}} \mathsf{C}_n\mathsf{P}\hat\pi^\mu_{n-1} \xrightarrow{\text{sampling}} \hat\pi^\mu_n := \mathsf{S}^N\mathsf{C}_n\mathsf{P}\hat\pi^\mu_{n-1}.$$

To see that this algorithm corresponds to the "optimal" SIR particle filter (cf. [19]), note that sampling from the measure $\mathsf{C}_n\mathsf{P}\hat\pi^\mu_{n-1}$, where $\hat\pi^\mu_{n-1} = \sum_{i=1}^N \delta_{X_{n-1}(i)}$, can be implemented as follows. Define two random variables X and Z with joint distribution

$$M(Z \in dz, X \in dx) := \frac{\hat\pi^\mu_{n-1}(dz)\, p(z, x)\, \varphi(dx)\, g(x, Y_n)}{\int \hat\pi^\mu_{n-1}(dz)\, p(z, x)\, \varphi(dx)\, g(x, Y_n)},$$

[3] Here we assume that the samples $X(1), \dots, X(N)$ generated by $\mathsf{S}^N\rho$ are the same as the samples generated from S^N_ρ, which is why we speak of identity between $\mathsf{C}_n\mathsf{S}^N\rho$ and $\mathsf{S}^N_\rho\mathsf{C}_n\rho$.

and note that $M(X \in \cdot) = \mathsf{C}_n\mathsf{P}\hat{\pi}^\mu_{n-1}$. *To sample* $\tilde{X} \sim M(X \in \cdot)$ *we can do the following:*

1. *sample* $\tilde{Z} \sim M(Z \in dz) = \frac{\hat{\pi}^\mu_{n-1}(dz) \int p(z,x)\,\varphi(dx)\,g(x,Y_n)}{\int \hat{\pi}^\mu_{n-1}(dz)\,p(z,x)\,\varphi(dx)\,g(x,Y_n)} = \sum_{i=1}^N W^\star_n(i)\,\delta_{X_{n-1}(i)}(dz)$;

2. *sample* $\tilde{X} \sim M(X \in dx | Z = \tilde{Z}) = \frac{p(\tilde{Z},x)\,\varphi(dx)\,g(x,Y_n)}{\int p(\tilde{Z},x)\,\varphi(dx)\,g(x,Y_n)}$;

where we have defined the "optimal" weights

$$W^\star_n(i) := \frac{\int p(X_{n-1}(i),x')\,\psi(dx)\,g(x',Y_n)}{\sum_{i=1}^N \int p(X_{n-1}(i),x')\,\psi(dx)\,g(x',Y_n)}.$$

Even if we were able to sample from the weighted measure $\mathsf{C}_n\mathsf{P}\hat{\pi}^\mu_{n-1}$ *as described above, this would still not resolve the curse of dimensionality in the filtering context. Indeed, the error between* $\pi^\mu_1 = \mathsf{F}_1\mu$ *and* $\hat{\pi}^\mu_1 = \hat{\mathsf{F}}_1\mu$ *would be dimension-free, namely,*

$$\|\pi^\mu_1 - \hat{\pi}^\mu_1\| = \|\mathsf{C}_1\mathsf{P}\mu - \mathsf{S}^N\mathsf{C}_1\mathsf{P}\mu\| \leq \frac{1}{\sqrt{N}},$$

but the error between $\pi^\mu_2 = \mathsf{F}_2\pi^\mu_1$ *and* $\hat{\pi}^\mu_2 = \hat{\mathsf{F}}_2\hat{\pi}^\mu_1$ *would again exhibit exponential dependence on the dimension due to the sampling performed in the first time step. The curse of dimensionality would therefore still arise due to the recursive nature of the filtering problem (see also [46]).*

Chapter 4

Block particle filter

This chapter is to develop the main framework of local particle filters that can overcome the curse of dimensionality. This is achieved by providing a detailed analysis of the block particle filter that we presently introduce. Emphasis is given to the decay of correlations property, which is seen to be the key to establish spatially uniform error bounds, thus representing the spatial counterpart of filter stability. The material here presented builds on the ideas introduced at the end of the previous chapter, and it is instrumental for the next chapter. This chapter is based on the paper [40].

4.1 Filtering models in high dimension

In order to investigate filtering problems in high dimension in a systematic way, we presently introduce a class of high-dimensional filtering models that will provide the basic framework to be investigated throughout this chapter and the next one. In these models, the state (X_n, Y_n) at each time n is a random field $(X_n^v, Y_n^v)_{v\in V}$ indexed by a (finite) undirected graph $G = (V, E)$. The graph G describes the spatial degrees of freedom of the model, and the underlying dynamics and observations are local with respect to the graph structure in a sense to be made precise below. The dimension of the model should be interpreted as the cardinality of the vertex set V, which is typically assumed to be large. Our aim is to develop quantitative results that are, under appropriate assumptions, independent of the dimension $\operatorname{card} V$.

We now define the hidden Markov model $(X_n, Y_n)_{n\geq 0}$ to be considered in the sequel (we will adopt throughout the basic setting and notation introduced in Section 3.1). The state spaces $\mathbb{X}$ and $\mathbb{Y}$ of X_n and Y_n, and the reference measures ψ and φ of the transition densities p and g, respectively, are of product form

$$\mathbb{X} = \prod_{v\in V} \mathbb{X}^v, \qquad \mathbb{Y} = \prod_{v\in V} \mathbb{Y}^v, \qquad \psi = \bigotimes_{v\in V} \psi^v, \qquad \varphi = \bigotimes_{v\in V} \varphi^v,$$

where ψ^v and φ^v are reference measures on the Polish spaces $\mathbb{X}^v$ and $\mathbb{Y}^v$, respectively. The transition densities p and g are given by

$$p(x, z) = \prod_{v\in V} p^v(x, z^v), \qquad g(x, y) = \prod_{v\in V} g^v(x^v, y^v),$$

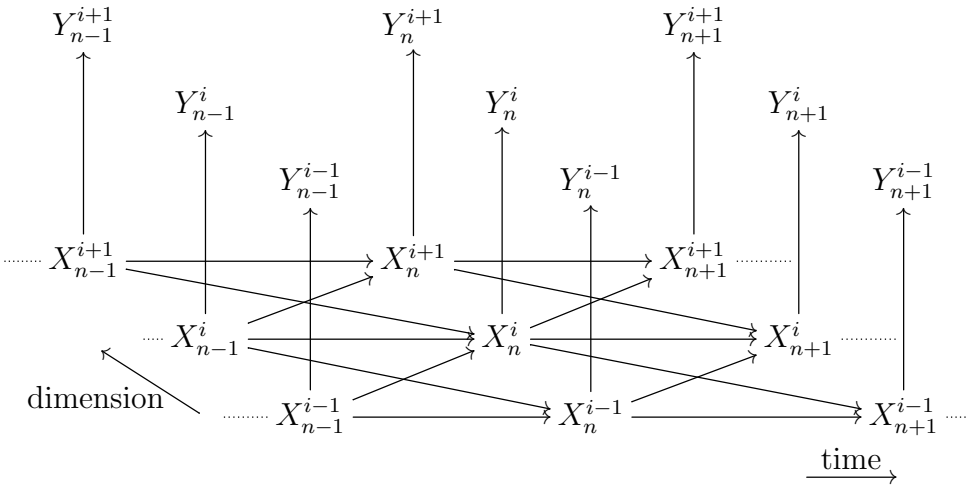

Figure 4.1: Dependency graph of a high-dimensional filtering model of the type considered in this chapter.

where $p^v : \mathbb{X} \times \mathbb{X}^v \to \bar{\mathbb{R}}_+$ and $g^v : \mathbb{X}^v \times \mathbb{Y}^v \to \bar{\mathbb{R}}_+$ are transition densities with respect to the reference measures ψ^v and φ^v, respectively.

The spatial graph G is endowed with its natural distance d (that is, $d(v, v')$ is the length of the shortest path in G between $v, v' \in V$). Let us fix throughout a neighborhood size $r \in \mathbb{N}$, and define for each vertex $v \in V$ the r-neighborhood

$$N(v) := \{v' \in V : d(v, v') \leq r\}.$$

We will assume that the dynamics of the underlying process $(X_n)_{n\geq 0}$ is *local* in the sense that $p^v(x, z^v)$ depends on $x^{N(v)}$ only (we write $x^J = (x^j)_{j\in J}$ for $J \subseteq V$):

$$p^v(x, z^v) = p^v(\tilde{x}, z^v) \quad \text{whenever} \quad x^{N(v)} = \tilde{x}^{N(v)}.$$

That is, the conditional distribution of X^v_n given $X_0, \ldots, X_{n-1}$ depends on $X^{N(v)}_{n-1}$ only. Similarly, by construction, the observations are local in that the conditional distribution of Y^v_n given X_n depends on X^v_n only. This dependence structure is illustrated in Figure 4.1 (in the simplest case of a linear graph G with $r = 1$).

Markov models of the form introduced above appear in the literature under various names, such as locally interacting Markov chains or probabilistic cellular automata [16, 35]. Such models arise naturally in numerous complex and large-scale applications, including percolation models of disease spread or forest fires, freeway traffic flow models, probabilistic models on networks and large-scale queueing systems, and various biological, ecological and neural models. Moreover, local Markov processes of this type arise naturally from finite-difference approximation of stochastic partial differential equations, and are therefore in principle applicable to a diverse set of data assimilation problems that arise in areas such as weather forecasting, oceanography, and geophysics (cf. Section 4.4.4). While more general models are certainly of substantial interest, the model defined above is prototypical of a broad range of high-dimensional data assimilation problems and provides a basic setting for the investigation of filtering problems in high dimension.

4.2 Decay of correlations and localization

As was explained in Section 3.3.2, the SIR particle filter is not well suited to address high-dimensional filtering models: the approximation error generally grows exponentially in the model dimension $\operatorname{card} V$. However, in the trivial case when the signal dynamics does not couple neighbors, that is, $r = 1$ (this is the analogue of the trivial model introduced in Section 3.3.2), we know an algorithm that can overcome the curse of dimensionality: we can simply run the SIR particle filter independently to each of the chains constituting the model. Clearly, in this way the error bound pertaining each single marginal of the model (that is, each chain) is, by construction, independent of the model dimension.

When the signal dynamics couples neighbors ($r > 1$), however, the law of the model at each spatial location is no longer independent. Nonetheless, large-scale interacting systems can exhibit an approximate version of independence among coordinates: this is the *decay of correlations* phenomenon that has been particularly well studied in statistical mechanics (see, e.g., [27]). Informally speaking, while the states (X_n^v, Y_n^v) and (X_n^w, Y_n^w) at two sites $v, w \in V$ are probably quite strongly correlated when v and w are close together, one might expect that (X_n^v, Y_n^v) and (X_n^w, Y_n^w) are nearly independent when v and w are far apart as measured with respect to the natural distance d in the graph G. The idea is that due to the decay of correlations, also in the care $r > 1$ the model can be "locally low-dimensional", in the sense that the conditional distribution of each coordinate only needs to be updated by observations in a neighborhood whose size is independent of the ambient dimension. Roughly speaking, the "local dimension" of the model is the number of coordinates in a ball whose radius is the correlation length of the filtering distribution.

As seen in Section 3.3.1, the sampling step added to the original filter recursion is the key to exploit algorithmically filter stability and get particle filters (i.e., the SIR particle filter) that yield time-uniform error bounds. In this chapter we will demonstrate that proper forms of *localization* of the filter recursion can be used to exploit algorithmically the decay of correlations property and to design *local* particle filters that yield error bounds that are uniform both in time and in the model dimension.

A speculative back-of-the-envelope computation explains how this might work. Due to the decay of correlations, the conditional distribution of the site X_n^v given the new observation Y_n should not depend significantly on observations Y_n^w at sites w distant from v. Suppose we can develop a local particle filtering algorithm that at each site v only uses observations in a local neighborhood K of v to update the filtering distribution. As we have now restricted to observations in K, the sampling error at each site will be exponential only in $\operatorname{card} K$ rather than in the full dimension $\operatorname{card} V$. On the other hand, the truncation to observations in K is only approximate: the decay of correlations property suggests that the bias introduced by this truncation should decay exponentially in $\operatorname{diam} K$. Therefore,

$$\text{error} = \text{bias} + \text{variance} \approx e^{-\operatorname{diam} K} + \frac{e^{\operatorname{card} K}}{\sqrt{N}}.$$

If the size of the neighborhoods K is chosen so as to optimize the error, then the resulting algorithm is evidently consistent (with a slower convergence rate than the standard $1/\sqrt{N}$ Monte Carlo rate: this is likely unavoidable in high dimension) with an error bound that is independent of the model dimension $\operatorname{card} V$.

4.3 Block particle filter

In this chapter we will investigate in detail the simplest possible local particle filtering algorithm that can exploit decay of correlations properties of the underlying filtering model, the *block particle filter*. While this algorithm possesses some inherent limitations (see Section 4.4.3 below), it is the simplest local algorithm both mathematically and computationally, and therefore provides an ideal starting point for the investigation of particle filters in high dimension.

To define the block particle filtering algorithm, we begin by introducing a partition $\mathcal{K}$ of the vertex set V into nonoverlapping blocks: that is, we have

$$V = \bigcup_{K\in\mathcal{K}} K, \qquad K \cap K' = \varnothing \text{ for } K \neq K',\ K, K' \in \mathcal{K}.$$

We now define the blocking operator

$$\mathsf{B}\rho := \bigotimes_{K\in\mathcal{K}} \mathsf{B}^K \rho,$$

where for any measure ρ on $\mathbb{X} = \bigotimes_{v\in V} \mathbb{X}^v$ and $J \subseteq V$ we denote by $\mathsf{B}^J\rho$ the marginal of ρ on $\bigotimes_{v\in J} \mathbb{X}^v$. The random field described by the measure $\mathsf{B}\rho$ on $\mathbb{X}$ is independent across different blocks defined by the partition $\mathcal{K}$, while the marginal on each block agrees with the original measure ρ. The block particle filter inserts an additional blocking step into the SIR particle filter recursion: that is,

$$\hat\pi_0^\mu = \mu, \qquad \hat\pi_n^\mu = \hat{\mathsf{F}}_n \hat\pi_{n-1}^\mu \quad (n \geq 1),$$

where $\hat{\mathsf{F}}_n := \mathsf{C}_n\mathsf{B}\mathsf{S}^N\mathsf{P}$ consists of four steps

$$\begin{aligned}
\hat\pi_{n-1} &\xrightarrow{\text{prediction}} \mathsf{P}\hat\pi_{n-1} &&\xrightarrow{\text{sampling}} \mathsf{S}^N\mathsf{P}\hat\pi_{n-1} \\
&\xrightarrow{\text{blocking}} \mathsf{B}\mathsf{S}^N\mathsf{P}\hat\pi_{n-1} &&\xrightarrow{\text{correction}} \hat\pi_n := \mathsf{C}_n\mathsf{B}\mathsf{S}^N\mathsf{P}\hat\pi_{n-1}.
\end{aligned}$$

The resulting algorithm is given in Figure 4.2. Figure 4.3 illustrates a typical iteration of the algorithm. In the special case $\mathcal{K} = \{V\}$, the block particle filter reduces to the SIR particle filter, so that the former is a strict generalization of the latter (we have therefore not introduced a separate notation for the SIR particle filter: in this chapter, the notation $\hat\pi_n^\mu$ always refers to the block particle filter).

The introduction of independent blocks allows to localize the algorithm, which will be crucial in the high-dimensional setting. We can immediately see this fact if we apply the block particle filter to the trivial model obtained with $r = 1$: choosing

Algorithm 3: Block particle filter

Data: Fix $n, N \geq 1$. Let the observations $Y_1, \ldots, Y_n$ be given.
Let $\hat{\pi}_0^\mu = \mu$;
for $k = 1, \ldots, n$ **do**
  Sample i.i.d. $Z_{k-1}(i)$, $i = 1, \ldots, N$ from the distribution $\hat{\pi}_{k-1}^\mu$;
  Sample $X_k^v(i) \sim p^v(Z_{k-1}(i), \cdot)\, d\psi^v$, $i = 1, \ldots, N$, $v \in V$;
  Compute $W_k^K(i) = \frac{\prod_{v \in K} g^v(X_k^v(i), Y_k^v)}{\sum_{\ell=1}^N \prod_{v \in K} g^v(X_k^v(\ell), Y_k^v)}$, $i = 1, \ldots, N$, $K \in \mathcal{K}$;
  Let $\hat{\pi}_k^\mu = \bigotimes_{K \in \mathcal{K}} \sum_{i=1}^N W_k^K(i)\, \delta_{X_k^K(i)}$;
Compute the approximate filter $\pi_n^\mu f \approx \hat{\pi}_n^\mu f$.

Figure 4.2: The block particle filtering algorithm considered in this chapter.

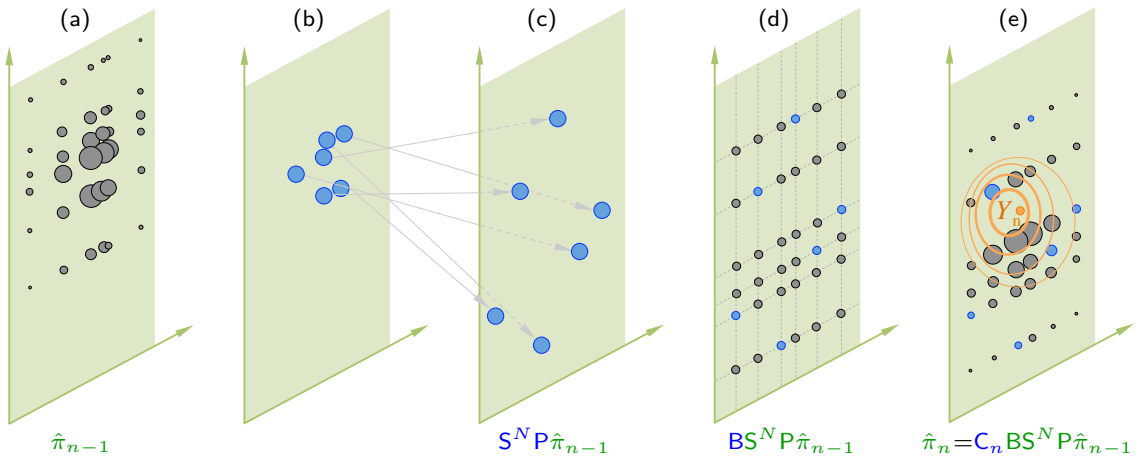

Figure 4.3: Representation of a single iteration of the block particle filter, with $\mathbb{X} = \mathbb{Y} = \mathbb{R}_+^2$ and $N = 6$. Each particle is represented by a ball, whose size is proportional to the weight of the particle. (a) Representation of $\hat{\pi}_{n-1}$. (b) Resampling step: N particles are sampled independently with replacement and weights are reset to $1/N$. (c) Particles are propagated forward using the underlying dynamics. (d) Blocking step: grey balls represent the "ghost" particles that are generated by shuffling the coordinates of the existing N particles (blue balls). (e) Particles are reweighed according to the likelihood of the new observation at time n (whose level sets are drawn in orange) yielding $\hat{\pi}_n$.

$\mathcal{K} = \bigcup_{v \in V} \{v\}$ the algorithm reduces to applying the SIR particle filter independently to each of the chains constituting the model; that is, we recover the original algorithm that motivated our discussion in the first place.

The rest of the chapter is devoted to showing that the localization procedure introduced by the blocking step can indeed overcome the curse of dimensionality even in the more realistic case of a coupled dynamics, $r > 1$ (proofs are provided in Appendix A). Note that in this case the blocking step introduces some bias in the algorithm, so that the estimates given by the block particle filter do not converge to the exact filter distributions as the number of particles N goes to infinity. However, the hope is that by introducing a small amount of bias in the algorithm, its variance can be reduced significantly.

In fact, it is immediately evident from inspection of the block particle filtering algorithm that only observations in block K are used by the algorithm to update the filtering distribution in block K. Therefore, following the heuristic ideas discussed in the Section 4.2, we expect that the sampling error of the algorithm is exponential in $\operatorname{card} K$ rather than in the model dimension $\operatorname{card} V$. To control the bias introduced by the blocking step, note that the blocking operator $\mathsf{B}\rho$ decouples the distribution ρ at the boundaries of the blocks. The decay of correlations property (if it can be established) should cause the influence of such a perturbation on the marginal distribution at a vertex $v \in K$ to decay exponentially in the distance from v to the boundary of the block K. Thus the back-of-the-envelope computation in Section 4.2 applies to the local error at "most" vertices, as the boundaries of the blocks only constitute a small fraction of the total number of vertices. On the other hand, the error will necessarily be larger for vertices closer to the block boundaries. This spatial inhomogeneity of the local error is an inherent limitation of the block particle filter that one might hope to alleviate by the development of more sophisticated local particle filters. We postpone further discussion of this point to Section 4.4.3.

Remark 4.1 (On distributed computing)**.** *By their nature, local particle filtering algorithms, such as the block particle filter here considered, are well suited to distributed computation: as the particles are updated locally in the spatial graph, this opens the possibility of implementing each local neighborhood on a separate processor. While this was not the original intention of the algorithms we propose, such properties could prove to be advantageous in their own right for the practical implementation of filtering algorithms in very large-scale systems.*

4.4 Main result: error bounds uniform in the dimension

Having introduced the block particle filtering algorithm, we now proceed to formulate the main result of this chapter (Theorem 4.2 below).

Recall that we have introduced the neighborhoods

$$N(v) := \{v' \in V : d(v, v') \leq r\}$$

above, where the neighborhood size r is fixed throughout this chapter (in our model, the state of vertex v depends only on the states of vertices in $N(v)$ in the previous time step). Given a set $J \subseteq V$, we denote the r-inner boundary of J as

$$\partial J := \{v \in J : N(v) \not\subseteq J\}$$

(that is, ∂J is the subset of vertices in J that can interact with vertices outside J in one step of the dynamics). We also define the following quantities:

$$\begin{aligned} |\mathcal{K}|_\infty &:= \max_{K\in\mathcal{K}} \operatorname{card} K, \\ \Delta &:= \max_{v\in V} \operatorname{card}\{v' \in V : d(v,v') \le r\}, \\ \Delta_{\mathcal{K}} &:= \max_{K\in\mathcal{K}} \operatorname{card}\{K' \in \mathcal{K} : d(K,K') \le r\}, \end{aligned}$$

where we define as usual $d(J,J') := \min_{v\in J}\min_{v'\in J'} d(v,v')$ for $J, J' \subseteq V$. Thus $|\mathcal{K}|_\infty$ is the maximal size of a block in $\mathcal{K}$, while Δ ($\Delta_{\mathcal{K}}$) is the maximal number of vertices (blocks) that interact with a single vertex (block) in one step of the dynamics. It should be emphasized that r, Δ and $\Delta_{\mathcal{K}}$ are *local* quantities that depend on the geometry but not on the size of the spatial graph G.

Finally, we introduce for $J \subseteq V$ the local distance

$$|||\rho - \rho'|||_J := \sup_{f\in\mathcal{X}^J : |f|\le 1} \sqrt{\mathbf{E}\,|\rho(f) - \rho'(f)|^2}$$

between random measures ρ, ρ' on $\mathbb{X}$, where $\mathcal{X}^J$ denotes the class of measurable functions $f : \mathbb{X} \to \bar{\mathbb{R}}$ such that $f(x) = f(\tilde{x})$ whenever $x^J = \tilde{x}^J$.

Theorem 4.2 (Block particle filter, main result)**.** *There exists a constant $0 < \varepsilon_0 < 1$, depending only on the local quantities Δ and $\Delta_{\mathcal{K}}$, such that the following holds.*

Suppose there exist $\varepsilon_0 < \varepsilon < 1$ and $0 < \kappa < 1$ such that

$$\varepsilon \le p^v(x, z^v) \le \varepsilon^{-1}, \qquad \kappa \le g^v(x^v, y^v) \le \kappa^{-1} \qquad \forall\, v \in V,\ x, z \in \mathbb{X},\ y \in \mathbb{Y}.$$

Then for every $n \ge 0$, $x \in \mathbb{X}$, $K \in \mathcal{K}$ and $J \subseteq K$ we have

$$|||\pi_n^x - \hat{\pi}_n^x|||_J \le \alpha \operatorname{card} J \left[e^{-\beta_1 d(J,\partial K)} + \frac{e^{\beta_2 |\mathcal{K}|_\infty}}{\sqrt{N}} \right],$$

where the constants $0 < \alpha, \beta_1, \beta_2 < \infty$ depend only on ε, κ, r, Δ, and $\Delta_{\mathcal{K}}$.

The key point of this result is that both the assumptions and the resulting error bound depend only on *local* quantities. In particular, the assumptions and error bound depend neither on time n nor on the model dimension $\operatorname{card} V$.

Remark 4.3 (On the assumptions of Theorem 4.2)**.** *A threshold requirement of the form $\varepsilon > \varepsilon_0$ is essential in order to obtain the decay of correlations property: the decay of correlations can fail if $\varepsilon > 0$ is too small (a phenomenon known as* phase transition *in statistical mechanics). Otherwise, the assumptions of Theorem 4.2 are comparable*

to assumptions commonly imposed in the literature to obtain error bounds for the SIR particle filter [8, 15] and possess similar limitations. We postpone a discussion of these issues to Section 4.4.1 below. Let us also note that explicit expressions for the constants in Theorem 4.2 can be read off from the proofs; however, we do not believe that our methods are sufficiently sharp to yield practical quantitative results.

Remark 4.4 (Dependence on observations)**.** *The particle filter $\hat{\pi}_n^\mu$ depends both on the random samples that are drawn in the algorithm and on the random sequence of the observations. However, the randomness of the observations plays no role in our proofs. One can therefore interpret the expectation in the definition of $|||\cdot|||_J$ as being taken only with respect to the random sampling mechanism in the block particle filter, and the bound of Theorem 4.2 as holding uniformly with respect to the observation sequence.*

Remark 4.5 (Initial measure)**.** *In Theorem 4.2 we have considered π_n^x and $\hat{\pi}_n^x$ with a non-random initial condition $x \in \mathbb{X}$. This is a choice of convenience: the proof of Theorem 4.2 yields the same conclusion for more general initial conditions that satisfy a suitable decay of correlations property. On the other hand, the stability property of the filter (e.g., Corollary A.5) ensures that π_n^μ forgets its initial condition μ exponentially fast uniformly in the dimension, so there is little loss of generality in choosing a computationally convenient initial condition.*

To provide a concrete illustration of Theorem 4.2, we consider in the remainder of this section the example where the spatial graph G is a square lattice, that is,

$$V = \{-d, \ldots, d\}^q \qquad (d, q \in \mathbb{N})$$

endowed with its natural edge structure. Note that in this case, the graph distance $d(v, v')$ is simply the ℓ_1-distance between the corresponding vectors of integers. To define the partition $\mathcal{K}$, we cover V by blocks of radius $b \in \mathbb{N}$: that is,

$$\mathcal{K} = \{(x + \{-b, \ldots, b\}^q) \cap V : x \in (2b+1)\mathbb{Z}^q\}.$$

We assume for simplicity in the sequel that $b \geq r$, and that $(2d+1)/(2b+1) \in \mathbb{N}$ is integer so that all $K \in \mathcal{K}$ are translates of $\{-b, \ldots, b\}^q$ (this slightly simplifies our arguments below but is not essential to our results). We can easily compute

$$|\mathcal{K}|_\infty = (2b+1)^q, \qquad \Delta \leq (2r+1)^q, \qquad \Delta_{\mathcal{K}} \leq 3^q.$$

Note that these local quantities do not depend on the size d of our lattice. In a data assimilation application one might have, for example, $q = 2$, $r = 1$, $d \sim 10^3$.

Consider the block $K = \{-b, \ldots, b\}^q$. Note that for $u = 0, \ldots, b - r$

$$\{v \in K : d(v, \partial K) > u\} = \{-(b-r-u), \ldots, b-r-u\}^q.$$

Fix $0 < \delta < 1$ and choose $u = \lfloor \delta(2b+1)/2q - r \rfloor$. Then

$$\frac{\operatorname{card}\{v \in K : d(v, \partial K) > u\}}{\operatorname{card} K} = \left(\frac{2(b-r-u)+1}{2b+1}\right)^q \geq 1 - \delta,$$

where we have used $1 - (1-\delta)^{1/q} \geq \delta/q$. The same conclusion evidently holds for every block $K \in \mathcal{K}$. Thus Theorem 4.2 gives the following corollary.

Corollary 4.6. *In the square lattice setting $V = \{-d, \dots, d\}^q$, there exists a constant $0 < \varepsilon_0 < 1$, depending only on r and q, such that the following holds.*

Suppose there exist $\varepsilon_0 < \varepsilon < 1$ and $0 < \kappa < 1$ such that

$$\varepsilon \le p^v(x, z^v) \le \varepsilon^{-1}, \qquad \kappa \le g^v(x^v, y^v) \le \kappa^{-1} \qquad \forall\, v \in V,\ x, z \in \mathbb{X},\ y \in \mathbb{Y}.$$

Then for every $x \in \mathbb{X}$, $n \ge 0$, and $0 < \delta < 1$ we have

$$\operatorname{card}\left\{v \in V : \left|\!\left|\!\left|\pi_n^x - \hat{\pi}_n^x\right|\!\right|\!\right|_v \le \alpha' e^{-\beta_1' \delta (2b+1)} + \alpha' \frac{e^{\beta_2'(2b+1)^q}}{\sqrt{N}}\right\} \ge (1-\delta) \operatorname{card} V,$$

where the constants $0 < \alpha', \beta_1', \beta_2' < \infty$ depend only on ε, κ, r, and q.

In particular, if we choose the block size $b = \lfloor \frac{1}{2}(4\beta_2')^{-1/q} \log^{1/q} N - \frac{1}{2} \rfloor$, then

$$\operatorname{card}\left\{v \in V : \left|\!\left|\!\left|\pi_n^x - \hat{\pi}_n^x\right|\!\right|\!\right|_v \le c_1 e^{-c_2 \delta \log^{1/q} N}\right\} \ge (1-\delta) \operatorname{card} V$$

and

$$\frac{1}{\operatorname{card} V} \sum_{v \in V} \left|\!\left|\!\left|\pi_n^x - \hat{\pi}_n^x\right|\!\right|\!\right|_v \le \frac{c_3}{\log^{1/q} N},$$

where the constants $0 < c_1, c_2, c_3 < \infty$ depend only on ε, κ, r, and q.

Corollary 4.6 makes precise the notion that a properly tuned block particle filter can avoid the curse of dimensionality: choosing the block size $b \sim \log^{1/q} N$, we obtain a local error that can be made arbitrarily small, uniformly both in time n and in the lattice size d, by choosing a sufficiently large sample size N. More precisely, we see that the local error at *most* locations (i.e., on an arbitrarily large fraction of the graph) is of order $e^{-c \log^{1/q} N}$, which is polynomial for $q = 1$ and subpolynomial otherwise. The bound for the average local error is similarly uniform in n and d, albeit with a very slow convergence rate. It appears that these results are chiefly limited by the spatial inhomogeneity that is inherent in the block particle filtering algorithm, as will be discussed in Section 4.4.3 below.

Remark 4.7. *We have stated the local error in Corollary 4.6 in terms of one-dimensional marginals $\left|\!\left|\!\left|\pi_n^x - \hat{\pi}_n^x\right|\!\right|\!\right|_v$ for simplicity; an analogous result can be obtained for marginals over cubes of any fixed size $\left|\!\left|\!\left|\pi_n^x - \hat{\pi}_n^x\right|\!\right|\!\right|_{v+\{-s,\dots,s\}^q}$.*

Remark 4.8. *Theorem 4.2 and Corollary 4.6 should be viewed as a theoretical proof of concept that it is possible, in principle, to design particle filters that avoid the curse of dimensionality. In practice, the slow rate $b \sim \log^{1/q} N$ suggests that the block size must typically be quite small (of order unity) for realistic values of the sample size N, which yields a large bias term in our bounds. We have nonetheless observed in simple simulations that the algorithm can work quite well even with the choice $b = 0$, so that the practical utility of the algorithm may not be fully captured by our mathematical results. Moreover, specific features of certain data assimilation applications, such as sparsity of observations, could make it possible to choose substantially larger blocks.*

A systematic investigation of the empirical performance of local particle filtering algorithms in applications is beyond the scope of our analysis, however. The practical implementation of local particle filters for data assimilation will likely require further advances in all mathematical, methodological and applied aspects of high-dimensional filtering.

In the next three sections we discussion the main aspects of Theorem 4.2.

4.4.1 Mixing assumptions and the ergodicity threshold

The basic assumption of Theorem 4.2 is that the local transition densities are bounded above and below:

$$\varepsilon \le p^v(x, z^v) \le \varepsilon^{-1}, \qquad \kappa \le g^v(x^v, y^v) \le \kappa^{-1}.$$

This is a local counterpart of the mixing assumptions that are routinely employed in the analysis of particle filters [8, 15]. The global mixing assumption $\varepsilon \le p(x, z) \le \varepsilon^{-1}$ would imply that the underlying Markov chain is strongly ergodic (in the sense that its transition kernel is a strict contraction with respect to the total variation distance, cf. Lemma 2.10) and is often used to establish the stability property of the filter (cf. Theorem 3.7). This is essential to obtain a time-uniform bound on the particle filter error, see Section 3.3.1 and Section 4.5.1 below. The local mixing assumption $\varepsilon \le p^v(x, z^v) \le \varepsilon^{-1}$ employed here should similarly be viewed as a local ergodicity assumption on the model.

It is well known that strong mixing assumptions of this type impose some constraints on the underlying model. In particular, strong mixing assumptions often require a compact state space: in a noncompact state space the likelihood ratio $p(x, z)/p(x', z)$ is typically unbounded as $|z| \to \infty$ (this is readily verified in linear Gaussian models, for example), while $\varepsilon \le p(x, z) \le \varepsilon^{-1}$ would imply that $p(x, z)/p(x', z)$ is uniformly bounded. Similarly, the assumptions of Theorem 4.2 will typically only hold in models whose local state spaces $\mathbb{X}^v$ and $\mathbb{Y}^v$ are compact. While qualitative results in this area have been obtained in much more general settings (cf. [52] and the references therein), it has proved to be more difficult to obtain quantitative results under assumptions weaker than strong mixing conditions: it remains an open problem, for example, to obtain quantitative time-uniform bounds under mild ergodicity assumptions even for the approximation error of the SIR particle filter. These technical issues are however unrelated to the problems that arise in high dimension, and we do not address them here.

On the other hand, there is a crucial assumption in Theorem 4.2 that does not arise in finite dimension. In classical results on particle filters, it is assumed that $\varepsilon \le p(x, z) \le \varepsilon^{-1}$ with $\varepsilon > 0$. For the local assumption $\varepsilon \le p^v(x, z^v) \le \varepsilon^{-1}$, however, it is not sufficient to assume that $\varepsilon > 0$; we must assume that $\varepsilon > \varepsilon_0$ for some strictly positive threshold $\varepsilon_0 > 0$. Some assumption of this form is absolutely essential in the high-dimensional setting. Unlike the global mixing assumption, the local mixing assumption is not in itself sufficient to ensure that the underlying model will remain ergodic as the dimension $\operatorname{card} V \to \infty$: the cumulative effect of the interactions can

create long-range correlations that break both ergodicity and any decay of correlations properties. Typically, the model is ergodic when the mixing constant ε is sufficiently large, but ergodicity breaks abruptly as ε drops below a threshold value ε_0. Such phenomena, called *phase transitions* in statistical mechanics, are very common in large-scale interacting systems: see [35, 16] for a number of examples. When the underlying model fails to exhibit ergodicity and decay of correlations, we lack the mechanism that we aim to exploit by developing local particle filters. Therefore, some assumption of the form $\varepsilon > \varepsilon_0$ is essential in Theorem 4.2 in order to ensure the presence of decay of correlations.

Unfortunately, the actual constant ε_0 that arises in the proof of Theorem 4.2 is almost certainly far from optimal. The Dobrushin machinery (Theorem 2.11) that forms the basis of our proof already does not yield sharp estimates of the phase transition point even in the simplest classical models of statistical mechanics. It is also far from clear whether the block particle filter should necessarily possess the same phase transition point as the underlying model: it may be that the algorithm only works in a strict subset of the regime in which the underlying model possesses the decay of correlations property. The mathematical tools used in this chapter are not sufficiently powerful to address much more delicate questions of this type. The practical relevance of Theorem 4.2 is therefore of a qualitative nature—we show that the block particle filter can beat the curse of dimensionality above a certain phase transition point—but should not be relied upon to provide quantitative guidance in specific situations. It remains of substantial interest to weaken the assumptions of Theorem 4.2 and to obtain sharper quantitative results; further progress in this direction will require the development of a more sophisticated probabilistic toolbox for the investigation of filtering problems in high dimension.

It should be noted that the problems investigated in this chapter are closely related to fundamental properties of conditional distributions. We have implicitly taken for granted that the filter will be stable when the underlying model is ergodic (and similarly for the decay of correlations property), but it is far from obvious that such properties are in fact preserved under conditioning on the observations. While the inheritance of ergodic properties under conditioning can be proved in a very general setting for models with finite-dimensional observations (see [52] and the references therein), we will see in Chapter 7 that there exist surprising examples in infinite dimension where the filter is non-ergodic even though the underlying model is ergodic and nondegenerate. Such probabilistic phenomena remain poorly understood. The threshold assumption $\varepsilon > \varepsilon_0$ rules out such issues in the setting of this chapter.

4.4.2 Ergodicity in space and time

The intuition behind the block particle filtering algorithm is that the localization controls the sampling error (as it replaces the model dimension $\operatorname{card} V$ by the block size $|\mathcal{K}|_\infty$), while the decay of correlations property of the model controls the localization error (as it ensures that the effect of the localization decreases in the distance to the block boundary). This intuition is clearly visible in the conclusion of Theorem 4.2. It is however not automatically the case that our model does indeed exhibit decay of

correlations: when there are strong interactions between the vertices, phase transitions can arise and the decay of correlations can fail much as for standard models in statistical mechanics [35], in which case we cannot expect to obtain dimension-free performance for the block particle filter. Such phenomena are ruled out in Theorem 4.2 by the assumption that $\varepsilon \le p^v \le \varepsilon^{-1}$ for $\varepsilon > \varepsilon_0$, which ensures that the interactions in our model are sufficiently weak.

It is notoriously challenging to obtain sharp quantitative results for interacting models, and it is unlikely that one could obtain realistic values for the constants in Theorem 4.2 at the level of generality considered here. More concerning, however, is that the weak interaction assumption of Theorem 4.2 is already unsatisfactory at the *qualitative* level, as decay of correlations in space and time are treated on the same footing: as $\varepsilon \to 1$, both the spatial and temporal correlations disappear. Note that there is no interaction between the vertices in the extreme case $\varepsilon = 1$; the assumption $\varepsilon > \varepsilon_0$ should be viewed as a perturbation of this situation (i.e., weak interactions). However, setting $\varepsilon = 1$ turns off not only the interaction between different vertices, but also the interaction between the same vertex at different times: in this setting the dynamics of the model become trivial. In contrast, one would expect that it is only the strength of the spatial interactions, and not the local dynamics, that is relevant for dimension-free errors, so that Theorem 4.2 places an unnatural restriction on our understanding of block particle filters.

It is therefore of interest to separate the temporal and spatial ergodicity assumptions, for example, by replacing the assumption $\varepsilon \le p^v(x, z^v) \le \varepsilon^{-1}$ by an assumption of the form $\varepsilon\, q^v(x^v, z^v) \le p^v(x, z^v) \le \varepsilon^{-1} q^v(x^v, z^v)$ that only controls the spatial interactions, where the transition density q^v describes the local dynamics at the vertex v in the absence of interactions. Rather than assuming $p^v(x, z^v) \approx 1$ as in Theorem 4.2, we would like to assume only that the spatial interactions are weak in the sense that $p^v(x, z^v) \approx q^v(x^v, z^v)$.

Overcoming this deficiency behind Theorem 4.2 requires the development of more refined comparison theorems than the Dobrushin comparison theorem that is used repeatedly for the results presented in this chapter (see Section 4.5.2 below). This new toolbox is of its own interest, and it will be the subject of Chapter 6. The analysis of the block particle filter on the basis of the new comparison theorems will yield Theorem 6.13, which improves qualitatively Theorem 4.2.

4.4.3 Local algorithms and spatial homogeneity

The major drawback of the block particle filtering algorithm is the spatial inhomogeneity of the bias. As was explained in Section 4.3, the block particle filter introduces errors at the block boundaries. We will increase the size of the blocks as the number of particles N increases, so that more points are distant from the block boundaries and therefore benefit from the decay of correlations. Nonetheless, points near the boundary will always be subject to larger errors, and we can only hope to implement the intuition of Section 4.2 to spatial locations that are strictly in the interior of the blocks.

The consequences of this inhomogeneity are manifested quantitatively in Corollary 4.6. Near the block boundaries, Theorem 4.2 gives a bound of order unity. By excluding a small fraction of spatial locations, however, we eliminate the block boundaries to retain an error of order $e^{-c \log^{1/q} N}$ at "most" spatial locations:

$$\operatorname{card} \left\{ v \in V : |||\pi_n^x - \hat{\pi}_n^x|||_v \lesssim e^{-c\delta \log^{1/q} N} \right\} \geq (1-\delta) \operatorname{card} V.$$

If, on the other hand, we compute the spatial average of the error, we obtain an exceedingly slow convergence rate that is much worse than the "typical" rate:

$$\frac{1}{\operatorname{card} V} \sum_{v \in V} |||\pi_n^x - \hat{\pi}_n^x|||_v \lesssim \frac{1}{\log^{1/q} N}.$$

Note that the block boundaries constitute a fraction $\sim 1/b$ of spatial locations, where b is the block size; therefore, as $b \sim \log^{1/q} N$ in Corollary 4.6, we see that the error at the block boundaries dominates our bound on the average error.

The behavior of the errors described above seems to be an inherent limitation of the block particle filtering algorithm. It is therefore of significant interest to explore the possibility that one could develop alternative local particle filtering algorithms that are spatially homogeneous. Conceptually, as explained in Section 4.2, such an algorithm should update the filtering distribution at each site v using sites in a centered neighborhood $N_b(v) := \{v' \in V : d(v, v') \leq b\}$; the decay of correlations should then yield a bias that decays exponentially in b. In this case, we would expect to obtain a spatially uniform error bound of the form

$$\sup_{v \in V} |||\pi_n^x - \hat{\pi}_n^x|||_v \lesssim e^{-c \log^{1/q} N}$$

for the optimized neighborhood size $b \sim \log^{1/q} N$. Whether it is in fact possible to design a local particle filtering algorithm that attains such a uniform error bound is still an open question. Chapter 5 is devoted to discussing one possible idea that could be of interest in this setting.

4.4.4 High-dimensional models in data assimilation

The basic model that we have introduced in Section 4.1 is prototypical of many data assimilation problems and provides a particularly convenient mathematical setting for the investigation of filtering problems in high dimension. While such models could be directly relevant to many high-dimensional applications, there remains a substantial gap between relatively simple models of this form and realistic models used in the most complex applications, particularly in the geophysical, atmospheric and ocean sciences, that frequently consist of coupled systems of partial differential equations. The investigation of such complex problems, and the associated numerical, physical, and practical issues, is far beyond the scope of this thesis. We therefore restrict our discussion of such problems to a few brief comments.

In principle, discrete models as defined in Section 4.1 arise naturally as finite-difference approximations of stochastic partial differential equations with space-time

white noise forcing. As the resulting state spaces $\mathbb{X}^v$ are not compact, such systems cannot satisfy strong mixing assumptions (cf. Section 4.4.1), but this is likely a mathematical rather than a practical problem. More importantly, it is not clear whether the discretized models will be in the regime of decay of correlations (that is, above the phase transition point) even if the original continuum model possesses such properties. It is possible that this requirement would place constraints on the spatial and temporal discretization steps, in the spirit of the von Neumann stability criterion in numerical analysis. The physics of such problems could also impose constraints on the design of local particle filters; for example, it is suggested in [60, p. 4107] that discontinuities (such as might be introduced at the block boundaries in the block particle filtering algorithm) could generate spurious gravity waves in ocean models. Such numerical and practical issues are distinct from the fundamental problems in high dimension that we aim to address in this thesis, but can ultimately play an equally important role in complex applications.

Let us also note that models considered in the data assimilation literature are often deterministic partial differential equations without stochastic forcing; the only randomness in such models comes from the initial condition (cf. [34, 1]). In deterministic chaotic dynamical systems, it is impossible to obtain time-uniform approximations using classical particle filters as there is no dissipation mechanism for approximation errors (the filter cannot be stable in this case, cf. Section 4.5.1). This issue is not directly related to dimensionality issues in particle filters: such problems arise in every deterministic filtering problem. It is natural to regularize deterministic systems by adding dynamical noise to the model (there is an extensive literature on random perturbations of chaotic dynamics, see for example [6]); a similar observation has been made by practitioners in the context of ad-hoc filtering algorithms, cf. [34, section 5]. To our knowledge, a rigorous analysis of such ideas in the setting of particle filters has yet to be performed.

4.5 Outline of the proof: framework behind local particle filters

In this section we discuss the outline of the proof of the main result of this chapter, Theorem 4.2. While this discussion is tailored to the analysis of the block particle filter, the ideas here developed constitute the backbone of a more general framework that encompasses a new philosophy behind filtering in high dimension. The details of the proof of Theorem 4.2 will then be given in Appendix A.

4.5.1 Error decomposition

The goal of Theorem 4.2 is to bound the error between the filter π_n^μ and the block particle filter $\hat{\pi}_n^\mu$. Recall that both the filter (Section 3.1) and block particle filter (Section 4.3) are defined recursively:

$$\pi_n^\mu = \mathsf{F}_n \cdots \mathsf{F}_1 \mu, \qquad \hat{\pi}_n^\mu = \hat{\mathsf{F}}_n \cdots \hat{\mathsf{F}}_1 \mu,$$

where $\mathsf{F}_n := \mathsf{C}_n\mathsf{P}$ and $\hat{\mathsf{F}}_n := \mathsf{C}_n\mathsf{B}\mathsf{S}^N\mathsf{P}$. We introduce also the *block filter*

$$\tilde{\pi}_n^\mu := \tilde{\mathsf{F}}_\mathsf{n} \cdots \tilde{\mathsf{F}}_1 \mu$$

with $\tilde{\mathsf{F}}_n := \mathsf{C}_n\mathsf{B}\mathsf{P}$. By the triangle inequality, we have

$$|||\pi_n^\mu - \hat{\pi}_n^\mu|||_J \le \underbrace{|||\pi_n^\mu - \tilde{\pi}_n^\mu|||_J}_{\text{bias}} + \underbrace{|||\tilde{\pi}_n^\mu - \hat{\pi}_n^\mu|||_J}_{\text{variance}}.$$

The first term on the right-hand side quantifies the bias introduced by the projection on independent blocks, while the second term quantifies the error due to the variance of the random sampling in the algorithm. Each term will be bounded separately to obtain the two terms in the error bound of Theorem 4.2.

The challenges encountered in bounding the bias term (cf. Section 4.5.3) and the variance term (cf. Section 4.5.4) are quite different in nature. Nonetheless, both bounds are based on a basic scheme of proof that was invented in order to prove time-uniform bounds for the SIR particle filter [15, 8], see Section 3.3.1. We therefore begin by reviewing this general idea, which is based on a simple error decomposition.

Suppose for sake of illustration that we aim to bound directly the error between π_n^μ and $\hat{\pi}_n^\mu$. The basic idea is to write $\pi_n^\mu - \hat{\pi}_n^\mu$ as a telescoping sum:

$$\pi_n^\mu - \hat{\pi}_n^\mu = \sum_{s=1}^n \{\mathsf{F}_n \cdots \mathsf{F}_{s+1}\mathsf{F}_s\hat{\mathsf{F}}_{s-1} \cdots \hat{\mathsf{F}}_1\mu - \mathsf{F}_n \cdots \mathsf{F}_{s+1}\hat{\mathsf{F}}_s\hat{\mathsf{F}}_{s-1} \cdots \hat{\mathsf{F}}_1\mu\}.$$

By the triangle inequality,

$$|||\pi_n^\mu - \hat{\pi}_n^\mu||| \le \sum_{s=1}^n |||\mathsf{F}_n \cdots \mathsf{F}_{s+1}\mathsf{F}_s\hat{\pi}_{s-1}^\mu - \mathsf{F}_n \cdots \mathsf{F}_{s+1}\hat{\mathsf{F}}_s\hat{\pi}_{s-1}^\mu|||.$$

The term s in this sum could be interpreted as the contribution to the total error at time n due to the filter approximation made at time s.

The key insight is now that one can employ the *filter stability* property to control this sum uniformly in time. In its simplest form, this property can be proved in the following form (see Theorem 3.7): if $\varepsilon \le p(x,z) \le \varepsilon^{-1}$ for all $x, z \in \mathbb{X}$, then

$$|||\mathsf{F}_n \cdots \mathsf{F}_{s+1}\rho - \mathsf{F}_n \cdots \mathsf{F}_{s+1}\rho'||| \le 2\,\varepsilon^{-2}(1-\varepsilon^2)^{n-s}|||\rho - \rho'|||.$$

Thus, the filter forgets its initial condition at an exponential rate. However, this also means that past approximation errors are forgotten at an exponential rate: if we substitute the stability property in the above error decomposition, we obtain

$$|||\pi_n^\mu - \hat{\pi}_n^\mu||| \le \sum_{s=1}^n 2\,\varepsilon^{-2}(1-\varepsilon^2)^{n-s}|||\mathsf{F}_s\hat{\pi}_{s-1}^\mu - \hat{\mathsf{F}}_s\hat{\pi}_{s-1}^\mu||| \le 2\,\varepsilon^{-4}\sup_{n,\rho}|||\mathsf{F}_n\rho - \hat{\mathsf{F}}_n\rho|||.$$

Thus, if we can control the error $|||\mathsf{F}_n\rho - \hat{\mathsf{F}}_n\rho|||$ in a single time step, we obtain a time-uniform bound of the same order. In the case of the SIR particle filter, if $\kappa \le g(x,y) \le \kappa^{-1}$, we proved in Section 3.1 that $|||\mathsf{F}_n\rho - \hat{\mathsf{F}}_n\rho||| \le 2\kappa^{-2}/\sqrt{N}$.

The basic error decomposition discussed above allows us to separate the problem of obtaining time-uniform bounds into two parts: the one-step approximation error and the stability property. It is important to note, however, that both parts become problematic in high dimension. We have already seen (Section 3.3.2) that the one-step approximation error of the SIR particle filter is exponential in the model dimension; we will surmount this problem by working with the block particle filtering algorithm and performing a local analysis of the one-step error using the decay of correlations property (which must itself be established). On the other hand, the filter stability bound used above also becomes exponentially worse in high dimension: a local bound of the form $\varepsilon \le p^v(x,z^v) \le \varepsilon^{-1}$ only yields $\varepsilon^{\operatorname{card} V} \le p(x,z) \le \varepsilon^{-\operatorname{card} V}$, which is exponential in the model dimension $\operatorname{card} V$. To surmount this problem, we must develop a much more precise understanding of the filter stability property in high dimension, which proves to be closely related to the decay of correlations property. The development of these ingredients constitutes the bulk of the proof of Theorem 4.2.

4.5.2 Dobrushin comparison method

How can one control the approximation error of high-dimensional distributions? The basic idea that we aim to exploit, both algorithmically and mathematically, is that the decay of correlations property leads to a form of localization: the effect on the distribution in some spatial set J of a perturbation made in another set J' decays rapidly in the distance $d(J,J')$. Therefore, as long as we measure the error locally (in $\|\!|\cdot|\!\|_J$ rather than $\|\!|\cdot|\!\|$), one would hope to control the spatial accumulation of approximation errors much as we controlled the accumulation of approximation errors in time using the filter stability property.

The Dobrushin comparison theorem (Theorem 2.11) introduced in Section 2.4 is the tool that will allow us to characterize the crucial way in which the decay of correlations property enters the picture. In the current setting, a useful manifestation of the decay of correlations property is that the matrix D from the comparison theorem is such that D_{ij} decays exponentially in the distance $d(i,j)$. If this is in fact the case, then Theorem 2.11 yields, for example, $\|\rho-\tilde\rho\|_i \lesssim \sum_j e^{-d(i,j)} b_j$, where b_j measures the local error at site j between ρ and $\tilde\rho$ (in terms of the conditional distributions $\rho^j_\cdot$ and $\tilde\rho^j_\cdot$). The decay of correlations property therefore controls the accumulation of local errors much as one might expect.

Let us now explain how Theorem 2.11 will be applied in the filtering setting. For sake of illustration, consider the problem of obtaining a local filter stability bound: that is, we would like to bound $\|\pi_n^x - \pi_n^{\tilde x}\|_J$ for $x,\tilde x \in \mathbb{X}$ and $J \subseteq V$. It would seem natural to apply Theorem 2.11 directly with $I = V$, $\mathbb{S} = \mathbb{X}$, and $\rho = \pi_n^x$, $\tilde\rho = \pi_n^{\tilde x}$. This is not useful, however, as we do not know how to control the corresponding local quantities such as $\rho_z^v = \mathbf{P}^x(X_n^v \in \cdot\,|Y_1,\ldots,Y_n,X_n^{V\setminus\{v\}} = z^{V\setminus\{v\}})$.

Instead, define $I = \{0,\ldots,n\}\times V$ and $\mathbb{S} = \mathbb{X}^{n+1}$, and let

$$\rho = \mathbf{P}^x(X_0,\ldots,X_n \in \cdot\,|Y_1,\ldots,Y_n),$$
$$\tilde\rho = \mathbf{P}^{\tilde x}(X_0,\ldots,X_n \in \cdot\,|Y_1,\ldots,Y_n).$$

As

$$\|\pi_n^x - \pi_n^{\tilde{x}}\|_J = \|\rho - \tilde{\rho}\|_{\{n\}\times J},$$

we can now apply Theorem 2.11 to the *smoothing* distributions $\rho, \tilde{\rho}$. Unlike the filters $\pi_n^x, \pi_n^{\tilde{x}}$, however, ρ and ρ' are Markov random fields on I (cf. Figure 4.1), so that the conditional distributions $\rho_z^{k,v}$ and $\tilde{\rho}_z^{k,v}$ can be easily computed and controlled in terms of the local densities $p^v(x, z^v)$ and $g^v(x^v, y^v)$. For example, as

$$\rho(A) \propto \int \mathbf{1}_A(x, x_1, \ldots, x_n) \prod_{k=1}^{n} \prod_{v\in V} p^v(x_{k-1}, x_k^v)\, g^v(x_k^v, Y_k^v)\, \psi^v(dx_k^v),$$

and as $p^v(x_{k-1}, x_k^v)$ depends only on x_{k-1}^w for $d(w, v) \le r$, we obtain

$$\rho_z^{k,v}(B) \propto \int \mathbf{1}_B(z_k^v)\, p^v(z_{k-1}, z_k^v)\, g^v(z_k^v, Y_k^v) \prod_{w\in N(v)} p^w(z_k, z_{k+1}^w)\, \psi^v(dz_k^v)$$

for $0 < k < n$ and $v \in V$ (the proportionality is up to a normalization factor). We will repeatedly exploit expressions of this type to obtain explicit bounds on the quantities C_{ij} and b_j that appear in Theorem 2.11. It should be emphasized that $\rho_z^{k,v}$ is a genuinely local quantity: the product inside the integral contains at most $\operatorname{card} N(v) \le \Delta$ terms. We will consequently be able to use Theorem 2.11 to obtain bounds that do not depend on the model dimension $\operatorname{card} V$.

Remark 4.9. *In the language of statistical mechanics, we exploit the fact that the smoothing distribution* $\mathbf{P}^x(X_0, \ldots, X_n \in \cdot\,|Y_1, \ldots, Y_n)$ *is a* Gibbs measure *[27] on the space-time index set* $I = \{0, \ldots, n\} \times V$*. Similar insight has proved to be fruitful in the ergodic theory of large-scale interacting Markov chains, cf. [35].*

4.5.3 Bounding the bias: decay of correlations

To bound the bias $\|\pi_n^x - \tilde{\pi}_n^x\|_J$, we follow the basic error decomposition scheme described above: that is,

$$\|\pi_n^x - \tilde{\pi}_n^x\|_J \le \sum_{s=1}^{n} \|\mathsf{F}_n \cdots \mathsf{F}_{s+1}\mathsf{F}_s\tilde{\pi}_{s-1}^x - \mathsf{F}_n \cdots \mathsf{F}_{s+1}\tilde{\mathsf{F}}_s\tilde{\pi}_{s-1}^x\|_J.$$

To implement our program, we must now obtain suitable local bounds on the stability of the filter and on the one-step approximation error. Both these problems will be approached by application of the Dobrushin comparison theorem.

In its most basic form, one can prove a filter stability property of the following type: provided $\varepsilon > \varepsilon_0$, there exists $\beta > 0$ (depending only on Δ and r) such that

$$\|\mathsf{F}_n \cdots \mathsf{F}_{s+1}\mu - \mathsf{F}_n \cdots \mathsf{F}_{s+1}\nu\|_J \le 4 \operatorname{card} J\, e^{-\beta(n-s)}$$

for any probability measures μ, ν on $\mathbb{X}$ and $J \subseteq V$, $n \ge 0$ (cf. Corollary A.5). This bound is evidently dimension-free, unlike the crude filter stability bound described in Section 4.5.1. Nonetheless, this filter stability bound would yield a trivial result

when substituted in the error decomposition, as it does not provide any control in terms of the distance between μ and ν (and therefore in terms of the one-step error). Instead, we will prove in Section A.1 the local stability bound

$$\|\mathsf{F}_n \cdots \mathsf{F}_{s+1}\mu - \mathsf{F}_n \cdots \mathsf{F}_{s+1}\nu\|_J \le 2e^{-\beta(n-s)} \sum_{v \in J} \max_{v' \in V} e^{-\beta d(v,v')} D_{v'}(\mu,\nu),$$

where $D_{v'}(\mu,\nu)$ is a suitable measure of the local error between μ and ν at site v' that arises naturally from the Dobrushin comparison theorem (see Proposition A.2 for precise expressions). This filter stability bound is genuinely local: the stability on the spatial set $J \subseteq V$ depends predominantly on the local distance of the initial conditions near J (that is, the spatial accumulation of errors is mitigated). This localization comes at a price, however; the local filter stability bound holds only if the initial condition μ satisfies *a priori* a decay of correlations property.

Once the local filter stability bound is substituted in the error decomposition, it remains to prove a bound on the one-step error $D_v(\mathsf{F}_s\tilde{\pi}^x_{s-1}, \tilde{\mathsf{F}}_s\tilde{\pi}^x_{s-1})$ with respect to the local distance prescribed by the filter stability bound. This will be done in Section A.2: we will show that for a constant C that depends only on Δ, r, ε,

$$D_v(\mathsf{F}_s\mu, \tilde{\mathsf{F}}_s\mu) \le C\, e^{-\beta d(v,\partial K)}$$

for every $K \in \mathcal{K}$ and $v \in K$, provided again that μ satisfies *a priori* a decay of correlations property. This is precisely what we expect: as B only introduces errors at the block boundaries, the decay of correlations should ensure that the error at site v decays exponentially in the distance to the nearest block boundary. The Dobrushin comparison theorem allows to make this intuition precise.

The decay of correlations property evidently plays a dual role in our setting: it controls the approximation error of the block filter, which is the basic principle behind the block particle filtering algorithm; at the same time, it mitigates the spatial accumulation of approximation errors, which is essential for proving dimension-free bounds. In order to apply the above bounds, the key step that remains is to prove that the appropriate decay of correlations property does in fact hold, uniformly in time, for the block filter $\tilde{\pi}^x_n$. The latter will be shown in Section A.3 by iterating a one-step decay of correlations bound that is obtained once again using the Dobrushin comparison theorem. We conclude by putting together all these ingredients in Section A.4 to obtain a bound on the bias of the form

$$\|\pi^x_n - \tilde{\pi}^x_n\|_J \le C \operatorname{card} J\, e^{-\beta d(J,\partial K)}$$

for $J \subseteq K$ (Theorem A.12). This proves the first half of Theorem 4.2 (note that, as the bias does not depend on the random sampling in the block particle filtering algorithm, we can trivially replace $\|\pi^x_n - \tilde{\pi}^x_n\|_J$ by $|||\pi^x_n - \tilde{\pi}^x_n|||_J$ in this bound).

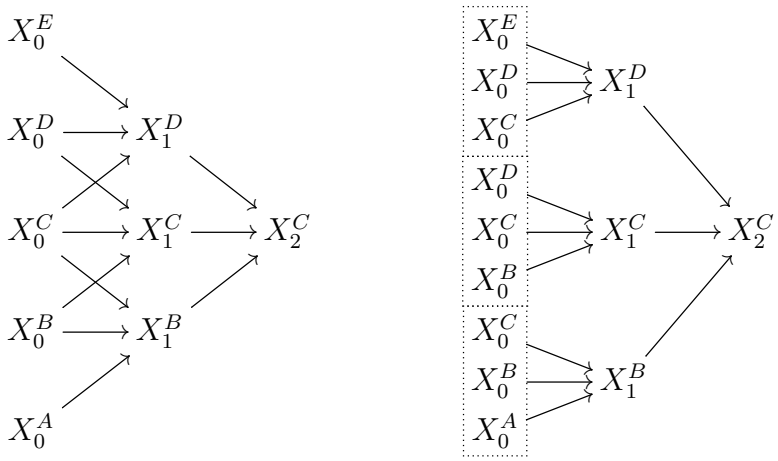

Figure 4.4: For a linear spatial graph G partitioned into blocks A–E (with $r = 1$), the dependencies between the blocks at subsequent times are illustrated here. The left dependency graph represents $\mathsf{B}^C\mathsf{P}^2\mu$, the right graph represents $\mathsf{B}^C\mathsf{PBP}\mu$. The blocking operation unravels the original graph into a tree by introducing independent duplicates (dotted boxes) of blocks in the previous time step.

4.5.4 Bounding the variance: the computation tree

To bound the variance term $\|\!|\!|\tilde{\pi}_n^x - \hat{\pi}_n^x|\!|\!|_J$, we once again start from the basic error decomposition

$$|\!|\!|\tilde{\pi}_n^x - \hat{\pi}_n^x|\!|\!|_J \le \sum_{s=1}^{n} |\!|\!|\tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_{s+1}\tilde{\mathsf{F}}_s\hat{\pi}_{s-1}^x - \tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_{s+1}\hat{\mathsf{F}}_s\hat{\pi}_{s-1}^x|\!|\!|_J.$$

The difficulties encountered in controlling this expression are quite different in nature, however, than what was needed to control the bias term.

Dimension-free bounds on the bias exploit decay of correlations: the core difficulty is to obtain local control of the error inside the blocks. The variance term, on the other hand, will already grow exponentially in the size of the blocks due to the exponential dependence of the sampling error on the dimension of the observations. There is therefore no need bound the error on a finer scale than a single block. This makes the analysis of the variance much less delicate than controlling the bias, and it is indeed not difficult to obtain a variance bound of the right order on a finite time horizon (but growing exponentially in time n).

The chief difficulty in controlling the variance is to obtain a time-uniform bound. Note that, in the error decomposition for the variance term, it is not stability of the filter π_n^μ that enters the picture but rather stability of the block filter $\tilde{\pi}_n^\mu$. Unlike the filter, however, which has by construction an interpretation as the marginal of a smoothing distribution, the block filter is defined by a recursive algorithm and not as a conditional expectation. It is therefore not entirely obvious how one could adapt the approach outlined in Section 4.5.2 to this setting.

The key idea that will be used to establish stability is that the block filter can nonetheless be viewed as the marginal of a suitably defined Markov random field,

just like the filter can be viewed as the marginal of a smoothing distribution. This random field, however, lives on a much larger index set than the original model. The basic idea behind the construction is illustrated in Figure 4.4 (disregarding the observations for simplicity of exposition). When we apply the transition operator P, each block interacts with its $\Delta_{\mathcal{K}}$ neighbors in the previous time step. However, if we subsequently apply the blocking operator B, then each block is replaced by an independent copy. This could be modelled equivalently by introducing independent duplicates of the blocks in the previous time step, and having each block interact with its own set of duplicates. This unravels the original dependency graph into a tree. By iterating this process, we can express the block filter as the marginal of a Markov random field defined on a tree that contains many independent duplicates of each block. We call this construction the *computation tree* in analogy with a similar notion that arises in the analysis of belief propagation algorithms [50].

With this construction in place, we can now obtain a stability bound for the block filter by applying the Dobrushin comparison theorem to the computation tree. This will be done in Section A.5 to obtain a bound of the following form: provided $\varepsilon > \varepsilon_0$, there exist $\beta, \beta' > 0$ (depending only on $\Delta, \Delta_{\mathcal{K}}, r$) such that

$$\max_{K\in\mathcal{K}} \|\tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_{s+1}\mu - \tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_{s+1}\nu\|_K \le e^{\beta'|\mathcal{K}|_\infty} e^{-\beta(n-s)} \max_{K\in\mathcal{K}} \|\mu^K - \nu^K\|$$

for any pair of initial conditions of product form $\mu = \bigotimes_{K\in\mathcal{K}} \mu^K$, $\nu = \bigotimes_{K\in\mathcal{K}} \nu^K$ (cf. Corollary A.16). Combining this bound with the error decomposition, we obtain in Section A.6 a time-uniform bound on the variance term of the form

$$\max_{K\in\mathcal{K}} |||\tilde{\pi}_n^x - \hat{\pi}_n^x|||_K \le C\,\frac{e^{\beta'|\mathcal{K}|_\infty}}{\sqrt{N}},$$

where we bound the one-step error in the same spirit as the computation for the SIR particle filter in Section 3.1 (however, a more involved argument is needed here to surmount the fact that the block filter stability bound is given in a total variation norm rather than the weaker norm $|||\cdot|||_K$). Thus Theorem 4.2 is proved.

Remark 4.10 (Alternative error decomposition)**.** *The reason we must consider stability of the block filter is that we have first split the error into the bias $|||\pi_n^x - \tilde{\pi}_n^x|||_J$ and variance $|||\tilde{\pi}_n^x - \hat{\pi}_n^x|||_J$ parts, and then applied the error decomposition to each term separately. One might hope to circumvent the problem by applying the error decomposition directly to the total error $|||\pi_n^x - \hat{\pi}_n^x|||_J$ as was illustrated in Section 4.5.1, and then splitting the one-step error terms in this bound into bias and variance parts:*

$$\begin{aligned}
&|||\mathsf{F}_n \cdots \mathsf{F}_{s+1}\mathsf{F}_s\hat{\pi}_{s-1}^\mu - \mathsf{F}_n \cdots \mathsf{F}_{s+1}\hat{\mathsf{F}}_s\hat{\pi}_{s-1}^\mu|||_J \\
&\quad\le |||\mathsf{F}_n \cdots \mathsf{F}_{s+1}\mathsf{F}_s\hat{\pi}_{s-1}^\mu - \mathsf{F}_n \cdots \mathsf{F}_{s+1}\tilde{\mathsf{F}}_s\hat{\pi}_{s-1}^\mu|||_J \\
&\qquad + |||\mathsf{F}_n \cdots \mathsf{F}_{s+1}\tilde{\mathsf{F}}_s\hat{\pi}_{s-1}^\mu - \mathsf{F}_n \cdots \mathsf{F}_{s+1}\hat{\mathsf{F}}_s\hat{\pi}_{s-1}^\mu|||_J.
\end{aligned}$$

In this case, only stability of the filter is needed to control the error accumulation.

Unfortunately, using this simpler approach it is impossible to obtain a nontrivial bound on the bias. Indeed, to control the one-step bias $D_v(\mathsf{F}_s\mu, \tilde{\mathsf{F}}_s\mu)$, it is essential that

μ satisfies a decay of correlations property. In Section 4.5.3, the error decomposition required us to obtain such a bound for $\mu = \tilde{\pi}^x_{s-1}$, and we showed that the block filter does indeed possess the requisite decay of correlations property. On the other hand, if we apply the error decomposition to the total error as above, one would have to obtain such a bound for $\mu = \hat{\pi}^x_{s-1}$. This is impossible, as $\hat{\pi}^x_{s-1}$ cannot possess a useful decay of correlations property within the blocks.

To see this, consider what happens when we apply the Dobrushin comparison theorem to an empirical measure $\rho = \frac{1}{N}\sum_{k=1}^N \delta_{X_k}$ with X_k i.i.d. $\sim \nu$. Suppose that $\nu = \bigotimes_{i\in I} \nu^i$ for some (nonatomic) measures ν^i: this is the extreme case where ν has no spatial correlations at all. Nonetheless, the empirical measure ρ will be maximally correlated: as each X^i_k is distinct with unit probability, we obtain $\rho^i_X = \delta_{X^i}$ for every $X \in \{X_1, \ldots, X_N\}$, so that $C_{ij} = 1$ for every $i \neq j$ in Theorem 2.11. We therefore see that sampling destroys decay of correlations (this is, in essence, the same phenomenon that causes the curse of dimensionality of particle filters). For this reason, it is essential to consider the bias and variance terms separately.

Chapter 5

Localized Gibbs sampler particle filter

This chapter is devoted to introducing a particle filter algorithm that implements a *spatially homogeneous* localization to overcome the curse of dimensionality, hence addressing the main drawback of the block particle filter analyzed in the previous chapter. While a complete analysis of this algorithm is still missing, we prove a one-step error bound for the bias term that illustrates the mechanism that can provide spatially homogenous approximations of the filter distribution. The goal of this chapter is also to show that the general idea of local particle filters is much broader than is suggested by the block particle filtering algorithm, and that the mathematical analysis developed in this thesis could in itself provide inspiration for further methodological developments. The material presented in this chapter is new and has not been submitted to publication yet.

Henceforth, we assume to work in the same setting introduced in Section 4.1.

5.1 Motivations

The block particle filter was introduced in Section 4.3 by localizing the SIR particle filter recursion $\hat{\pi}_n = \mathsf{C}_n\mathsf{S}^N\mathsf{P}\hat{\pi}_{n-1}$ to $\hat{\pi}_n = \mathsf{C}_n\mathsf{B}\mathsf{S}^N\mathsf{P}\hat{\pi}_{n-1}$, via the blocking operator B that projects probability measures to the product of their marginals over a fix partition $\mathcal{K}$ of the vertex set V.

At the heart of our main result (Theorem 4.2) lies the decay of correlations. In the proofs there we used an intuitive notion of decay of correlations of essentially the following form: a probability measure ρ on $\mathbb{X} = \prod_{v\in V}\mathbb{X}^v$ possesses the decay of correlations property if the effect on the conditional distribution $\rho(X^v \in \cdot\,|X^{V\setminus\{v\}} = x^{V\setminus\{v\}})$ of a perturbation to $x^{v'}$ decays exponentially in the distance $d(v, v')$ (cf. Sections 4.5.2 and A.1). The blocking operation evidently replaces these conditional distributions by

$$(\mathsf{B}\rho)(X^v \in A|X^{V\setminus\{v\}} = x^{V\setminus\{v\}}) = \rho(X^v \in A|X^{K\setminus\{v\}} = x^{K\setminus\{v\}})$$

for every $K \in \mathcal{K}$ and $v \in K$. Therefore, if ρ possesses the decay of correlations property, then the bias at site $v \in K$ incurred by the blocking operation decays exponentially in the distance between v and the boundary of K. On the other hand, the sampling error depends only on the dimension of the block, and not on the dimension of the entire system.

As we discussed in Section 4.4 (particularly in Section 4.4.3), the major drawback of the block particle filtering algorithm is precisely the spatial inhomogeneity of the bias, as the blocking introduces errors at the block boundaries: points near the boundaries will always be subject to larger errors. On the one hand, it is true that by optimizing the error bound in Theorem 4.2 we find that the size of the blocks increases as the number of particles N increases, so that more points are distant from the block boundaries and therefore benefit from the decay of correlations. On the other hand, our theory suggests that the size of the blocks typically increases slowly (logarithmically) with the number of particles (see Corollary 4.6 for a concrete example), so that we should not consider large blocks.

From this perspective, an approach to spatially homogeneous algorithms readily suggests itself: we should aim to replace B with another operator M that satisfies

$$(\mathsf{M}\rho)(X^v \in A | X^{V\setminus\{v\}} = x^{V\setminus\{v\}}) = \rho(X^v \in A | X^{N_b(v)\setminus\{v\}} = x^{N_b(v)\setminus\{v\}})$$

for every $v \in V$, where $N_b(v) := \{v' \in V : d(v,v') \le b\}$. The bias incurred by this operation decays exponentially in b uniformly for all v (it is therefore spatially homogeneous). On the other hand, as

$$(\mathsf{C}_n\mathsf{M}\mathsf{P}\rho)(X^v \in A | X^{V\setminus\{v\}} = x^{V\setminus\{v\}}) = \frac{\int \mathbf{1}_A(x^v)\, g^v(x^v, Y_n^v) \prod_{w\in N_b(v)} p^w(z, x^w)\, \rho(dz)\, \psi^v(dx^v)}{\int g^v(x^v, Y_n^v) \prod_{w\in N_b(v)} p^w(z, x^w)\, \rho(dz)\, \psi^v(dx^v)},$$

the sampling error incurred if we replace ρ by $\mathsf{S}^N\rho$ in this expression should only be exponential in $\operatorname{card} N_b(v)$ (which is $\sim b^q$ for the square lattice) rather than in the model dimension $\operatorname{card} V$. This suggests that the local particle filter defined by the recursion $\hat{\mathsf{F}}_n = \mathsf{S}^N\mathsf{C}_n\mathsf{M}\mathsf{P}$ should yield a spatially homogeneous algorithm in accordance with our intuition.

To implement this algorithm one needs to sample from the measure $\mathsf{C}_n\mathsf{M}\mathsf{P}\rho$, which we have defined only implicitly in terms of its conditional distributions. However, this is precisely the task to which Markov chain Monte Carlo (MCMC) methods are well suited. These methods sample from a probability measure by constructing a Markov chain that has the desired measure as its equilibrium distribution. In particular, the Gibbs sampler (Section 5.2 below) is a MCMC method that implements this paradigm by using transition kernels that are defined in terms of the conditional distributions of the desired measure.

One would therefore ostensibly obtain a spatially homogeneous local particle filtering algorithm that is recursive in time and that uses MCMC to sample the spatial degrees of freedom (regularization using M is still key to avoiding the curse of dimensionality, as replacing the sampling step in ordinary particle filters by an MCMC

method does not resolve the fundamental problem that we face in high dimension; see [3] for related discussion).

Conceptually, the idea introduced here is quite natural. The general idea of local particle filters is that one should introduce a spatial regularization step into the filtering recursion that enables local sampling. In the block particle filter, this regularization is provided by the blocking operation B that projects a probability measure on the class of measures that are independent across blocks. In the above algorithm, we aim to regularize instead by the operation M that projects a probability measure on the class of Markov random fields of order b. The fatal flaw in our reasoning is that the operator M that we have defined implicitly above does not exist: the truncated conditional distributions $\rho(X^v \in \cdot\,|X^{N_b(v)\setminus\{v\}} = x^{N_b(v)\setminus\{v\}})$ are typically not consistent, so there exists no single probability measure that satisfies our definition of $\mathsf{M}\rho$.

Nonetheless, the basic idea just discussed suggests a practical approach to approximating random fields by Markov random fields: we can substitute the above expression for $(\mathsf{C}_n\mathsf{M}\mathsf{P}\rho)(X^v \in \cdot\,|X^{V\setminus\{v\}})$ in a Gibbs sampler regardless of its inconsistency. The algorithm that we will introduce in this chapter, the *localized Gibbs sampler particle filter*, exactly implements this idea to yield spatially homogeneous estimates of the filter distribution.

While the analysis of the block particle filter relies heavily on the Dobrushin comparison theorem (Theorem 2.11), the analysis of the localized Gibbs sampler particle filter relies crucially on the one-sided Dobrushin comparison theorem (Theorem 2.12), which is needed to capture the directionality of time embedded in the definition of Gibbs samplers. Following the same bias/variance decomposition scheme adopted in Chapter 4, we will prove a spatially homogeneous one-step error bound for the bias of the localized Gibbs sampler particle filter (Theorem 5.4).

While this result is extremely encouraging, the analysis of the localized Gibbs sampler particle filter has proved to be much more challenging than the analysis of the block particle filter, and a complete picture is still lacking. With respect to the proof strategy followed in Chapter 4, the crucial difficulty lies in establishing a decay of correlations property for the approximate filter that is uniform in time. While we have strong reasons to believe that this property should hold, it seems that the Dobrushin comparison theorems are not adequate to capture it. A more delicate analysis is needed, with new tools to be developed.

5.2 Gibbs sampler

The backbone of the localized Gibbs sampler particle filter is the *Gibbs sampler*, a MCMC algorithm that samples from a high-dimensional distribution ρ on $\mathbb{X}$ by sampling iteratively from the low-dimensional distributions $\rho(X^v \in \cdot\,|X^{V\setminus\{v\}})$, $v \in V$. Henceforth in this chapter, label the elements of V as $\{v_1, \ldots, v_d\}$, where $d = \operatorname{card} V$, and introduce the notation $v_k : v_{k'} := \{v_k, v_{k+1} \ldots, v_{k'}\}$. The *systematic-scan* Gibbs sampler is the algorithm described in Figure 5.1.

Algorithm 4: Systematic-scan Gibbs sampler

Data: Fix $m \geq 1$, χ probability measure on $\mathbb{X}$.
Let $X_0 \sim \chi$;
for $\ell = 1, \dots, m$ **do**
  for $k = 1, \dots, d$ **do**
    Sample $X_\ell^{v_k} \sim \rho(X^{v_k} \in \cdot \,| X^{v_1:v_{k-1}} = X_\ell^{v_1:v_{k-1}}, X^{v_{k+1}:v_d} = X_{\ell-1}^{v_{k+1}:v_d})$;
Output: X_m.

Figure 5.1: Systematic-scan Gibbs sampler.

As described in Figure 5.1, in the ℓ-th round of the algorithm (that is needed to sample X_ℓ) each coordinate X_ℓ^v is cyclically obtained by sampling from the conditional distribution given all other coordinates $X_\ell^{V\setminus\{v\}}$. The cyclic sampling occurs *systematically*, following the ordering given by $v_1, \dots, v_d$[1]. Each round of the algorithm is usually referred to as a *sweep* of the algorithm.

The Gibbs sampler is a MCMC method that constructs a Markov chain $(X_n)_{n\geq 0}$ that admits ρ as its invariant measure (by construction ρ satisfies $\rho = \rho P$, where P is the transition kernel of the Markov chain). The main rationale is that if the Markov chain is quickly converging to equilibrium (*rapidly mixing*), then for large m we can reliably interpret X_m—the output of the algorithm in Figure 5.1—as a random variable whose distribution is close to ρ. We refer to [8] for an extensive treatment of MCMC methods in the context of filtering theory.

To facilitate the description of what follows, we introduce the (systematic-scan) Gibbs sampler sampling operator.

Given a probability measure ρ on $\mathbb{X}$ and $v \in V$, let G_ρ^v be the transition kernel defined as follows

$$G_\rho^v(x, A) := \int \rho(X^v \in d\omega^v | X^{V\setminus\{v\}} = x^{V\setminus\{v\}})\, \delta_{x^{V\setminus\{v\}}}(d\omega^{V\setminus\{v\}})\, \mathbf{1}_A(\omega).$$

Definition 5.1 (Gibbs sampler sampling operator)**.** *Let χ be a probability measure on $\mathbb{X}$. Define the* Gibbs sampler sampling operator $\mathsf{S}_\chi^{N,m}$ *as*

$$\mathsf{S}_\chi^{N,m}\rho = \mathsf{S}^N(\chi G_\rho^{v_1} \cdots G_\rho^{v_d})^m = \frac{1}{N}\sum_{i=1}^N \delta_{X(i)},$$

where $X(1), \dots, X(N)$ are i.i.d. samples—each obtained by running the algorithm described in Figure 5.1 with respect to the family of conditional distribution $(\rho(X^v \in \cdot\,|X^{V\setminus\{v\}}))_{v\in V}$—for m sweeps and with initial distribution χ. S^N is the sampling operator defined in Definition 2.16 (see Section 3.3 for a discussion on how to sample from Markov chains).

[1]Other sampling schemes can be considered, such as uniformly sampling d-times the elements of V at each round of the algorithm. This gives rise to the so-called *random-scan* Gibbs sampler.

Remark 5.2. *In the literature on Gibbs samplers the typical empirical measure that is considered has the following form*

$$\frac{1}{N}\sum_{k=0}^{N-1}\delta_{X_{n_0+k\ell}},$$

where $(X_n)_{n\geq 0}$ is the Markov chain generated by the algorithm in Figure 5.1, n_0 is the so-called burn-in period *that represents the amount of time it takes for the Markov chain to reach its invariant distribution (which is the measure we want to sample from), and ℓ is the period at which samples are taken into consideration (so to have samples that are close to being independent). On the other hand, in Definition 5.1 we consider samples that are independent by construction so to simplify the theoretical analysis of the algorithm.*

5.3 Gibbs sampler particle filter

As the block particle filter was introduced by localizing the SIR particle filter recursion, also the local algorithm that we analyze in this chapter comes as a localization of another particle filter—the *Gibbs sampler particle filter*—that we presently introduce.

Fix $N, m \geq 1$. We define the Gibbs sampler particle filter recursion as follows:

$$\hat{\pi}_0^\mu := \mu, \qquad \hat{\pi}_n^\mu := \mathsf{S}^{N,m}_{\hat{\pi}^\mu_{n-1}}\mathsf{C}_n\mathsf{P}\hat{\pi}^\mu_{n-1} \quad (n\geq 1),$$

where the recursion consists of three steps

$$\hat{\pi}^\mu_{n-1} \xrightarrow{\text{prediction}} \mathsf{P}\hat{\pi}^\mu_{n-1} \xrightarrow{\text{correction}} \mathsf{C}_n\mathsf{P}\hat{\pi}^\mu_{n-1} \xrightarrow{\substack{\text{MCMC}\\ \text{sampling}}} \hat{\pi}^\mu_n := \mathsf{S}^{N,m}_{\hat{\pi}^\mu_{n-1}}\mathsf{C}_n\mathsf{P}\hat{\pi}^\mu_{n-1}.$$

In Lemma 5.3 below we prove that under the usual mixing conditions considered in Chapter 4 ($\varepsilon \leq p^v \leq \varepsilon^{-1}$ for $0<\varepsilon<1$), as the number of sweeps m goes to infinity the Gibbs sampler particle filter recursion $\hat{\pi}^\mu_n = \mathsf{S}^{N,m}_{\hat{\pi}^\mu_{n-1}}\mathsf{C}_n\mathsf{P}\hat{\pi}^\mu_{n-1}$ converges to the "optimal" SIR particle filter recursion $\hat{\pi}^\mu_n = \mathsf{S}^N\mathsf{C}_n\mathsf{P}\hat{\pi}^\mu_{n-1}$ (cf. Remark 3.13). For this reason we have not introduced a separate notation for the Gibbs sampler particle filter and the SIR particle filter introduced in Chapter 3.

Presently, we illustrate one possible implementation of this algorithm. Let

$$\hat{\pi}^\mu_{n-1} = \frac{1}{N}\sum_{i=1}^{N}\delta_{X_{n-1}(i)},$$

where $X_{n-1}(1),\dots,X_{n-1}(N)$ are the samples coming from the $(n-1)$-th iteration of the Gibbs sampler sampling operator, that is, the samples coming from $\mathsf{S}^{N,m}_{\hat{\pi}^\mu_{n-2}}\mathsf{C}_{n-1}\mathsf{P}\hat{\pi}^\mu_{n-2}$. Recall that the Gibbs sampler samples iteratively from the conditional distributions of the measure it is applied to. While there are many ways to implement this sampling scheme (for instance, by using rejection-sampling to

directly sample from the conditional distributions, which are only needed to be known point-wise), we currently present a sampling procedure that takes place in multiple stages so to resemble the sampling scheme adopted for ordinary particle filters (with importance weights, see Chapter 3). This formulation will be functional to highlight, at least at a heuristic level, the reason why also the Gibbs sampler particle filter algorithm suffers from the curse of dimensionality, and it will readily suggest a way around the problem (see Section 5.4 below).

Notice that for any measure ρ on $\mathbb{X}$ we have

$$(\mathsf{C}_n\mathsf{P}\rho)(X^v \in A | X^{V\setminus\{v\}} = x^{V\setminus\{v\}}) = \frac{\int \rho(dz) \prod_{w\in V\setminus\{v\}} p^w(z, x^w)\, p^v(z,\omega)\, \psi^v(d\omega)\, g^v(\omega, Y_n^v)\, \mathbf{1}_A(\omega)}{\int \rho(dz) \prod_{w\in V\setminus\{v\}} p^w(z, x^w)\, p^v(z,\omega)\, \psi^v(d\omega)\, g^v(\omega, Y_n^v)}.$$

Hence, we can write

$$\begin{aligned}
&(\mathsf{C}_n\mathsf{P}\hat{\pi}^\mu_{n-1})(X^v \in A | X^{V\setminus\{v\}} = x^{V\setminus\{v\}}) \\
&\quad = \frac{\sum_{i=1}^N \prod_{w\in V\setminus\{v\}} p^w(X_{n-1}(i), x^w) \int p^v(X_{n-1}(i),\omega)\, \psi^v(d\omega)\, g^v(\omega, Y_n^v)\, \mathbf{1}_A(\omega)}{\sum_{i=1}^N \prod_{w\in V\setminus\{v\}} p^w(X_{n-1}(i), x^w) \int p^v(X_{n-1}(i),\omega)\, \psi^v(d\omega)\, g^v(\omega, Y_n^v)} \\
&\quad = \sum_{i=1}^N W^v_{n,x}(i)\, q^v_n(X_{n-1}(i), A),
\end{aligned}$$

where the weights are defined as

$$W^v_{n,x}(i) := \frac{Z^v_n(X_{n-1}(i)) \prod_{w\in V\setminus\{v\}} p^w(X_{n-1}(i), x^w)}{\sum_{i=1}^N Z^v_n(X_{n-1}(i)) \prod_{w\in V\setminus\{v\}} p^w(X_{n-1}(i), x^w)},$$

and q^v_n is a transition kernel from $\mathbb{X}$ to $\mathbb{X}^v$ defined as

$$q^v_n(z, A) := \frac{\int p^v(z,\omega)\, g^v(\omega, Y_n^v)\, \psi^v(d\omega)\, \mathbf{1}_A(\omega)}{Z^v_n(z)},$$

with

$$Z^v_n(z) := \int p^v(z,\omega)\, g^v(\omega, Y_n^v)\, \psi^v(d\omega).$$

As the weights are positive and $\sum_{i=1}^N W^v_{n,x}(i) = 1$ by construction, they can be interpreted as probabilities. So, sampling from $(\mathsf{C}_n\mathsf{P}\hat{\pi}^\mu_{n-1})(X^v \in \cdot\, | X^{V\setminus\{v\}} = x^{V\setminus\{v\}})$ can be achieved by first sampling J from the distribution $j \in \{1,\dots,N\} \to W^v_{n,x}(j)$, and then sampling from $q^v_n(X_{n-1}(J), \cdot\,)$ (note that this is a one-dimensional integral, and one can use one of the methods in [2] to sample from it, such as rejection-sampling). The resulting algorithm is given in Figure 5.2.[2]

[2]Note that the algorithm illustrated in Figure 5.2 differs a little from the one described in the main text, as for simplicity it is now assumed that $\hat{\pi}^\mu_0 = \frac{1}{N}\sum_{i=1}^N \delta_{X_0(i)}$, for $X_0(1),\dots,X_0(N) \sim \mu$ (if $\hat{\pi}^\mu_0 = \mu$ as in the main text, then the weights $W^v_{1,x}(j)$'s would be different). Moreover, note that more clever implementations of this algorithm can be considered, but this is beyond the scope of our current treatment.

Algorithm 5: Gibbs sampler particle filter

Data: Fix $n, m, N \geq 1$. Let the observations $Y_1, \ldots, Y_n$ be given.
Sample i.i.d. $X_0(i) \sim \mu$, $i = 1, \ldots, N$, and let $\hat{\pi}_0^\mu = \frac{1}{N} \sum_{i=1}^N \delta_{X_0(i)}$;
for $s = 1, \ldots, n$ **do**
  Sample i.i.d. $R_0(i)$, $i = 1, \ldots, N$, from the distribution $\hat{\pi}_{s-1}^\mu$;
  for $i = 1, \ldots, N$ **do**
    for $\ell = 1, \ldots, m$ **do**
      for $k = 1, \ldots, d$ **do**
        Let $R = (R_\ell^{v_1:v_{k-1}}(i), r, R_{\ell-1}^{v_{k+1}:v_d}(i))$, for any $r \in \mathbb{X}^{v_k}$;
        Sample J from the distribution $j \in \{1, \ldots, N\} \to W_{s,R}^{v_k}(j)$, with
        $W_{s,R}^{v_k}(j) = \frac{\prod_{w \in V \setminus \{v_k\}} p^w(X_{s-1}(j), R^w) \int p^{v_k}(X_{s-1}(j), \omega)\, \psi^{v_k}(d\omega)\, g^{v_k}(\omega, Y_s^{v_k})}{\sum_{j=1}^N \prod_{w \in V \setminus \{v_k\}} p^w(X_{s-1}(j), R^w) \int p^{v_k}(X_{s-1}(j), \omega) \psi^{v_k}(d\omega) g^{v_k}(\omega, Y_s^{v_k})}$;
        Sample $R_\ell^{v_k}(i) \sim q_s^{v_k}(X_{s-1}(J), d\omega) = \frac{p^{v_k}(X_{s-1}(J), \omega)\, g^{v_k}(\omega, Y_s^{v_k})\, \psi^{v_k}(d\omega)}{\int p^{v_k}(X_{s-1}(J), \omega)\, g^{v_k}(\omega, Y_s^{v_k})\, \psi^{v_k}(d\omega)}$;
  Let $X_s^v(i) = R_m^v(i)$, $i = 1, \ldots, N$, $v = v_1, \ldots v_d$, and $\hat{\pi}_s^\mu := \frac{1}{N} \sum_{i=1}^N \delta_{X_s(i)}$;
Compute the approximate filter $\pi_n^\mu f \approx \hat{\pi}_n^\mu f$.

Figure 5.2: Gibbs sampler particle filter.

5.4 Sample degeneracy with dimension

Also the Gibbs sampler particle filter runs into the curse of dimensionality. Ultimately, weight degeneracy occurs for the same reason why it occurs for the SIS algorithm (Section 3.2) and for the SIR algorithm (Section 3.3). To make this point, let us recall the definition of the weights (up to normalization factors) involved in these algorithms:

$$\begin{aligned}
\text{SIS particle filter} &\longrightarrow W_n(i) \propto \prod_{k=1}^n \prod_{v \in V} g^v(X_k^v(i), Y_k^v), \\
\text{SIR particle filter} &\longrightarrow W_n(i) \propto \prod_{v \in V} g^v(X_n^v(i), Y_n^v), \\
\text{Gibbs sampler particle filter} &\longrightarrow W_{n,x}^v(i) \propto Z_n^v(X_{n-1}(i)) \prod_{w \in V \setminus \{v\}} p^w(X_{n-1}(i), x^w),
\end{aligned}$$

(clearly, different algorithms involve different particles). Heuristically it is easy to see where the problem of weight degeneracy comes from: roughly speaking, weights get picked towards zero or infinity exponentially fast with the dimension $\operatorname{card} V$. In the SIS particle filter the problem appears both with time and space (see Section 3.2.1). In the SIR particle filter the problem is caused by the product of observation likelihoods (see Section 3.3.3), while in the Gibbs sampler particle filter the problem is caused by the product of transition likelihoods.

Proceeding in the same line of thoughts, it is easy to see why the block particle filter (see Section 4.3 for its definition) can overcome the curse of dimensionality.

Note, in fact, that the weights involved in this algorithm read

$$\text{Block particle filter} \quad \longrightarrow \quad W_k^K(i) \propto \prod_{v \in K} g^v(X_n^v(i), Y_n^v),$$

where the product of observation likelihoods is restricted only to the coordinates in the block $K \subseteq V$. Hence, the block particle filter samples at each coordinate v by using weights that are defined only through coordinates contained in the element K of the partition $\mathcal{K}$ such that $v \in K$. So, even if the dimensionality of the whole model ($\operatorname{card} V$) is increased, what matters for the sake of weight degeneracy is only the dimensionality of the blocks ($\operatorname{card} K$).

Following this intuition, a spatially-homogenous procedure to localize the Gibbs particle filter readily suggests itself. If in this algorithm we replace the measure $(\mathsf{C}_n\mathsf{P}\rho)(X^v \in A | X^{V\setminus\{v\}} = x^{V\setminus\{v\}})$ with the following measure

$$\frac{\int \rho(dz) \prod_{w \in N_b(v)\setminus\{v\}} p^w(z, x^w)\, p^v(z, \omega)\, \psi^v(d\omega)\, g^v(\omega, Y_n^v)\, \mathbf{1}_A(\omega)}{\int \rho(dz) \prod_{w \in N_b(v)\setminus\{v\}} p^w(z, x^w)\, p^v(z, \omega)\, \psi^v(d\omega)\, g^v(\omega, Y_n^v)}, \tag{5.1}$$

where the product over $w \in V$ is replaced with the product over $w \in N_b(v) := \{v' \in V : d(v, v') \leq b\}$, then the new algorithm—which we call *localized Gibbs sampler particle filter*—would yield weights of the following form:

Localized Gibbs sampler particle filter

$$\longrightarrow \quad W_{n,x}^v(i) \propto Z_n^v(X_{n-1}(i)) \prod_{w \in N_b(v)\setminus\{v\}} p^w(X_{n-1}(i), x^w).$$

That is, the new algorithm samples at each coordinate v by using weights that are defined only through coordinates contained in a ball of radius b centered at v. Thus, we obtain a spatially homogeneous way of localizing the sampling step, using the Gibbs sampler as a way of *constructing* a high-dimensional distribution from its conditional distributions (as discussed in Section 5.1, recall that this localization can not be described as $\hat{\pi}_n^\mu = \mathsf{S}^N\mathsf{C}_n\mathsf{M}\mathsf{P}\hat{\pi}_{n-1}^\mu$, since the measure $\mathsf{M}\mathsf{P}\hat{\pi}_{n-1}^\mu$ does not exists). The resulting algorithm is immediately given as in Figure 5.2 upon truncating the weights as we just mentioned.

5.5 Localized Gibbs sampler particle filter

We now introduce a more convenient description of the localized Gibbs sampler particle filter. For each probability measure ρ on $\mathbb{X}$ and each $n \geq 1$, $v \in V$, define the probability kernels $\eta_{n,\rho}^v$ and $\tilde{\eta}_{n,\rho}^v$ from $\mathbb{X}$ to $\mathbb{X}^v$ respectively as:

$$\eta_{n,\rho}^v(x, A) := \frac{\int \rho(dz) \prod_{w \in V\setminus\{v\}} p^w(z, x^w)\, p^v(z, \omega)\, \psi^v(d\omega)\, g^v(\omega, Y_n^v)\, \mathbf{1}_A(\omega)}{\int \rho(dz) \prod_{w \in V\setminus\{v\}} p^w(z, x^w)\, p^v(z, \omega)\, \psi^v(d\omega)\, g^v(\omega, Y_n^v)},$$

$$\tilde{\eta}_{n,\rho}^v(x, A) := \frac{\int \rho(dz) \prod_{w \in N_b(v)\setminus\{v\}} p^w(z, x^w)\, p^v(z, \omega)\, \psi^v(d\omega)\, g^v(\omega, Y_n^v)\, \mathbf{1}_A(\omega)}{\int \rho(dz) \prod_{w \in N_b(v)\setminus\{v\}} p^w(z, x^w)\, p^v(z, \omega)\, \psi^v(d\omega)\, g^v(\omega, Y_n^v)}.$$

It is easy to verify that

$$\begin{aligned}\eta^v_{n,\rho}(x,A) &= (\mathsf{C}_n\mathsf{P}\rho)(X^v \in A | X^{V\setminus\{v\}} = x^{V\setminus\{v\}}) \\ &= \mathbf{P}^\rho(X_1^v \in A | Y_1 = Y_n, X_1^{V\setminus\{v\}} = x^{V\setminus\{v\}}),\end{aligned}$$

while

$$\tilde{\eta}^v_{n,\rho}(x,A) = \mathbf{P}^\rho(X_1^v \in A | Y_1^{N_b(v)} = Y_n^{N_b(v)}, X^{N_b(v)\setminus\{v\}} = x^{N_b(v)\setminus\{v\}})$$

corresponds to the localized quantity (5.1). Let us also define the probability kernels $G^v_{n,\rho}$ and $\tilde{G}^v_{n,\rho}$ from $\mathbb{X}$ to $\mathbb{X}$ respectively as

$$\begin{aligned}G^v_{n,\rho}(x,A) &:= \int \eta^v_{n,\rho}(x,d\omega^v)\,\delta_{x^{V\setminus\{v\}}}(d\omega^{V\setminus\{v\}})\,\mathbf{1}_A(\omega), \\ \tilde{G}^v_{n,\rho}(x,A) &:= \int \tilde{\eta}^v_{n,\rho}(x,d\omega^v)\,\delta_{x^{V\setminus\{v\}}}(d\omega^{V\setminus\{v\}})\,\mathbf{1}_A(\omega),\end{aligned}$$

and the operators on probability measures

$$\begin{aligned}(\mathsf{G}^v_{n,\rho}\mu)f &:= \int \mu(dx)\,G^v_{n,\rho}(x,dx')\,f(x'), \\ (\tilde{\mathsf{G}}^v_{n,\rho}\mu)f &:= \int \mu(dx)\,\tilde{G}^v_{n,\rho}(x,dx')\,f(x').\end{aligned}$$

From the definition of the Gibbs sampler sampling operator (Definition 5.1) we can write

$$\mathsf{S}^{N,m}_\rho\mathsf{C}_n\mathsf{P}\rho = \mathsf{S}^N\rho(G^{v_1}_{n,\rho}\cdots G^{v_d}_{n,\rho})^m \equiv \mathsf{S}^N(\mathsf{G}^{v_d}_{n,\rho}\cdots\mathsf{G}^{v_1}_{n,\rho})^m\rho.$$

Therefore, the Gibbs sampler particle filter can be formulated as

$$\hat{\pi}^\mu_0 := \mu, \qquad\qquad \hat{\pi}^\mu_n = \mathsf{S}^N(\mathsf{G}^{v_d}_{n,\hat{\pi}^\mu_{n-1}}\cdots\mathsf{G}^{v_1}_{n,\hat{\pi}^\mu_{n-1}})^m\,\hat{\pi}^\mu_{n-1}.$$

At this point it is straightforward to describe the localization procedure previously discussed and to define the localized Gibbs sampler particle filter as

$$\hat{\pi}^\mu_0 := \mu, \qquad\qquad \hat{\pi}^\mu_n := \mathsf{S}^N(\tilde{\mathsf{G}}^{v_d}_{n,\hat{\pi}^\mu_{n-1}}\cdots\tilde{\mathsf{G}}^{v_1}_{n,\hat{\pi}^\mu_{n-1}})^m\,\hat{\pi}^\mu_{n-1}.$$

In the special case $b = \max_{v,v'\in V} d(v,v')$ the localized Gibbs sampler particle filter reduces to the Gibbs sampler particle filter, so that the former is a strict generalization of the latter (we have therefore not introduced a separate notation for the localized Gibbs sampler particle filter: in the remaining of this chapter, the notation $\hat{\pi}^\mu_n$ always refers to the localized Gibbs sampler particle filter).

Before moving to the analysis of the localized Gibbs sampler particle filter in the next section, we now prove that under the mixing conditions $\varepsilon \le p^v \le \varepsilon^{-1}$ for $0 < \varepsilon < 1$ the Gibbs sampler particle filter recursion $\hat{\pi}^\mu_n = \mathsf{S}^{N,m}_{\hat{\pi}^\mu_{n-1}}\mathsf{C}_n\mathsf{P}\hat{\pi}^\mu_{n-1}$ converges in the limit of infinitely many sweeps ($m \to \infty$) to the "optimal" SIR particle filter recursion $\hat{\pi}^\mu_n = \mathsf{S}^N\mathsf{C}_n\mathsf{P}\hat{\pi}^\mu_{n-1}$ (cf. Remark 3.13). Recall that $\mathsf{F}_n := \mathsf{C}_n\mathsf{P}$.

Lemma 5.3 (Convergence of Gibbs sampler particle filter)**.** *Suppose there exists* $0 < \varepsilon < 1$ *such that*

$$\varepsilon \le p^v(x, z^v) \le \varepsilon^{-1}, \qquad \forall\, v \in V,\ x, z \in \mathbb{X}.$$

Then, for each probability measures ρ, μ *on* $\mathbb{X}$ *and each* $n \ge 1$ *we have*

$$\lim_{\ell \to \infty} \|\mathsf{F}_n \rho - \mu(G_{n,\rho}^{v_1} \cdots G_{n,\rho}^{v_d})^\ell\| = 0.$$

Proof. From the local mixing conditions for each p^v we get the following minorization condition for each $\eta_{n,\rho}^v$,

$$\eta_{n,\rho}^v(x, A) \ge \varepsilon^{2(d-1)} \chi_n^v(A),$$

where

$$\chi_n^v(A) := \frac{\int \rho(dz)\, p^v(z, \omega)\, \psi^v(d\omega)\, g^v(\omega, Y_n^v)\, \mathbf{1}_A(\omega)}{\int \rho(dz)\, p^v(z, \omega)\, \psi^v(d\omega)\, g^v(\omega, Y_n^v)}.$$

Hence, we also get the following minorization condition for an entire sweep of the Gibbs sampler,

$$(G_{n,\rho}^{v_1} \cdots G_{n,\rho}^{v_d})(x, A) \ge \varepsilon^{2d(d-1)} \chi(A),$$

where $\chi(A) := \int \bigotimes_{v \in V} \chi_n^v(dz^v)\, \mathbf{1}_A(z)$. As by construction for each $v \in V$ the kernel $G_{n,\rho}^v$ leaves invariant the measure $\mathsf{F}_n \rho$, that is,

$$(\mathsf{F}_n \rho) G_{n,\rho}^v = \mathsf{F}_n \rho,$$

then by Lemma 2.10 we have

$$\begin{aligned} \|\mathsf{F}_n \rho - \mu(G_{n,\rho}^{v_1} \cdots G_{n,\rho}^{v_d})^\ell\| &= \|(\mathsf{F}_n \rho)(G_{n,\rho}^{v_1} \cdots G_{n,\rho}^{v_d})^\ell - \mu(G_{n,\rho}^{v_1} \cdots G_{n,\rho}^{v_d})^\ell\| \\ &\le (1 - \varepsilon^{2d(d-1)})^\ell \|\mathsf{F}_n \rho - \mu\|. \end{aligned}$$

□

5.6 Main result: spatially-homogeneous error bound

Ultimately, we would like to mimic the result of Theorem 4.2 for the localized Gibbs sampler particle filter, and to prove a bound for $\|\!|\pi_n^\mu - \hat{\pi}_n^\mu|\!\|_J$ that is uniform both in time (n) and in the model dimension ($\operatorname{card} V$), and that is spatially-homogeneous in $J \subseteq V$. Although at the time being we do not have such a result, Theorem 5.4 below represents an encouraging first step towards establishing it.

Following the strategy pursued in Chapter 4, we define the *approximate Gibbs sampler filter* as

$$\tilde{\pi}_0^\mu := \mu, \qquad \tilde{\pi}_n^\mu := \tilde{\mathsf{F}}_n\, \tilde{\pi}_{n-1}^\mu \equiv (\tilde{\mathsf{G}}_{n,\tilde{\pi}_{n-1}^\mu}^{v_d} \cdots \tilde{\mathsf{G}}_{n,\tilde{\pi}_{n-1}^\mu}^{v_1})^m\, \tilde{\pi}_{n-1}^\mu,$$

and we consider the following error decomposition (cf. Section 4.5.1)

$$|||\pi_n^\mu - \hat{\pi}_n^\mu|||_J \le \underbrace{|||\pi_n^\mu - \tilde{\pi}_n^\mu|||_J}_{\text{bias}} + \underbrace{|||\tilde{\pi}_n^\mu - \hat{\pi}_n^\mu|||_J}_{\text{variance}}.$$

Recall the following definitions from Chapter 4 and Appendix A. For any probability measure μ on $\mathbb{X}$ and $x, z \in \mathbb{X}$, $v, v' \in V$, $\beta > 0$, let

$$\begin{aligned}
\Delta &:= \max_{v \in V} \operatorname{card}\{v' \in V : d(v, v') \le r\}, \\
\mu_x^v(dz^v) &:= \mu(dz^v | x^{V \setminus \{v\}}), \\
\mu_{x,z}^v(A) &:= \frac{\int \mathbf{1}_A(x^v) \prod_{w \in N(v)} p^w(x, z^w)\, \mu_x^v(dx^v)}{\int \prod_{w \in N(v)} p^w(x, z^w)\, \mu_x^v(dx^v)}, \\
C_{vv'}^\mu &:= \frac{1}{2} \sup_{z \in \mathbb{X}} \sup_{x, \tilde{x} \in \mathbb{X} : x^{V \setminus \{v'\}} = \tilde{x}^{V \setminus \{v'\}}} \|\mu_{x,z}^v - \mu_{\tilde{x},z}^v\|, \\
\operatorname{Corr}(\mu, \beta) &:= \max_{v \in V} \sum_{v' \in V} e^{\beta d(v,v')} C_{vv'}^\mu.
\end{aligned}$$

The following result provides a spatially homogeneous one-step error bound for the bias term of the localized Gibbs sampler particle filter.

Theorem 5.4 (Localized Gibbs sampler particle filter, one-step error for the bias)**.** *There exists a constant $0 < \varepsilon_0 < 1$ depending only on the local quantity Δ such that the following holds.*

Suppose there exists $\varepsilon_0 < \varepsilon < 1$ such that

$$\varepsilon \le p^v(x, z^v) \le \varepsilon^{-1} \qquad \text{for all } v \in V,\ x, z \in \mathbb{X},$$

and let ρ be a probability measure on $\mathbb{X}$ such that

$$\operatorname{Corr}(\rho, \beta) \le \frac{1}{4},$$

where $\beta = \frac{1}{r+1} \log \frac{1}{8\Delta(1-\varepsilon^2)}$. Then, for each $n \ge 1$ and $J \subseteq V$ we have

$$\|\mathsf{F}_n \rho - \tilde{\mathsf{F}}_n \rho\|_J \le \alpha \operatorname{card} J\, e^{-\gamma \min\{b, m\}}.$$

where the constants $0 < \alpha, \gamma < \infty$ depend only on ε, r, and Δ.

We refer to Appendix B for the proof of Theorem 5.4. While in the case of the block particle filter the key insight to perform the analysis is that both filter and approximate block filter can be thought of as Gibbs measures on properly-defined graphs (see Section 4.5.2 and Remark 4.9), in the present case the key insight is that both filter and approximate Gibbs sampler filter can be thought of as Gibbs samplers.

In fact, as by construction for each $v \in V$ the kernel $G_{n,\rho}^v$ leaves the measure $\mathsf{F}_n \rho$ invariant (that is, $(\mathsf{F}_n \rho) G_{n,\rho}^v = \mathsf{F}_n \rho$), then we can express the filter recursion as m sweeps of a Gibbs sampler, namely,

$$\mathsf{F}_n \rho = (\mathsf{F}_n \rho)(G_{n,\rho}^{v_1} \cdots G_{n,\rho}^{v_d})^m.$$

On the other hand, the approximate Gibbs sampler filter recursion is defined as

$$\tilde{\mathsf{F}}_n\rho := \rho\,(\tilde{G}_{n,\rho}^{v_1}\cdots\tilde{G}_{n,\rho}^{v_d})^m.$$

The key idea to bound $\|\mathsf{F}_n\rho - \tilde{\mathsf{F}}_n\rho\|_J$ is then to use the one-sided Dobrushin comparison theorem (Theorem 2.12) to capture the one-sidedness that is embedded in the Gibbs samplers $\mathsf{F}_n\rho$ and $\tilde{\mathsf{F}}_n\rho$.

Remark 5.5 (On running the algorithm)**.** *The one-step error bound in Theorem 5.4 suggests that in practice we only need to run the localized Gibbs sampler particle filter for a number of sweeps (m) that is of the same order as the radius (b) at which we implement the localization. From the analysis of the block particle filter (see Section 4.4) we know that the optimal b increases quite slowly with the number of particles N ($b \sim \log^{1/q} N$ for the square lattice $V = \{-d,\ldots,d\}^q$), which suggests that each iteration of the algorithm does not need to be run for many sweeps.*

5.7 Where things stand

Theorem 5.4 yields a bound on the one-step error $\|\mathsf{F}_n\rho - \tilde{\mathsf{F}}_n\rho\|_J$ under a certain assumption on the decay of correlations for the measure ρ. In order to use this result within the general error decomposition scheme pursued in Section 4.5.3 to bound the bias term $\|\!|\pi_n^\mu - \tilde{\pi}_n^\mu|\!\|_J$, we need to prove that the appropriate decay of correlations property does in fact hold, uniformly in time, for the approximate filter $\tilde{\pi}_n^\mu$. That is, we would like to prove that

$$\sup_{n\geq 0} \mathrm{Corr}(\tilde{\pi}_n^\mu, \beta) \leq c < 1,$$

where c is an absolute constant which does not depend on the ambient dimension.

In the case of the block particle filter we can show this property by iterating a one-step decay of correlations bound that is obtained using the Dobrushin comparison theorem (see Section A.3). In the case of the localized Gibbs sampler particle filter, however, the situation is more involved as we need to control the way the decay of correlations is propagated in each iteration of the Gibbs samplers. While we have strong reasons to believe that the decay of correlations of the approximate filter should hold uniformly in time, at the time being we have been not successful in establishing the required behavior using the Dobrushin comparison method, and new mathematical tools seem to be needed.

To see why we expect the decay of correlations property to hold, consider the case of the filter recursion. While the Dobrushin comparison theorem can be used to bound the quantity $\mathrm{Corr}(\mathsf{F}_n\rho, \beta)$ by making an assumption on $\mathrm{Corr}(\rho, \beta)$ (as done in Section A.3, see Proposition A.9 in particular), it seems not possible to use the same machinery to bound $\mathrm{Corr}(\rho(G_{n,\rho}^{v_1}\cdots G_{n,\rho}^{v_d})^\ell, \beta)$, for any given finite ℓ, without making higher-order assumptions on the decay of correlations of ρ, although we know that $\lim_{\ell\to\infty} \rho(G_{n,\rho}^{v_1}\cdots G_{n,\rho}^{v_d})^\ell = \mathsf{F}_n\rho$ as seen in Lemma 5.3.

The approach that we have presented to bound the bias of the localized Gibbs sampler particle filter is taken from the analysis of the block particle filter given in Chapter 4, and it is based on the recursive property of the filter. On the other hand, the improved analysis of the block particle filter that will be given in the next chapter (Section 6.4) is based on another strategy that allows to directly use the Dobrushin comparison theorem on properly-defined space-time Gibbs measures, without considering the filter recursion. This new method yields a shorter proof for the bound of the bias term that does not involve controlling the decay of correlations quantity $\mathrm{Corr}(\tilde{\pi}_n^{\mu}, \beta)$. This approach relies on the ability to express both filter and approximate filter as the marginal of properly-defined Markov random fields, where the natural interaction range of the system (recall that the models introduced in Section 4.1 have an interaction of range r) can be recovered through the interaction neighborhood of the field (cf. the discussion on the Dobrushin comparison method in Section 4.5.2).

The problem in implementing this approach in the case of the localized Gibbs sampler particle filter lies in the fact that this algorithm is defined in terms of a recursion that does not seem to admit an intrinsic probabilistic interpretation that can allow to recover the natural interaction range of the model. Ultimately, the problem is that this algorithm is defined in terms of conditional probabilities that do not have a local structure (see Section 5.5). In fact, by definition, $\tilde{\eta}_{n,\rho}^{v}(x, A)$ depends on x^w whenever $d(v, w) \leq b$. Therefore, even if we can interpret the measure $\tilde{\pi}_n^{\mu}$ as the marginal of a properly-defined space-time Gibbs measure, this measure is a Markov random field with interaction neighborhood size b, which does not correspond to the intrinsic neighborhood size r.

Chapter 6

Comparison theorems for Gibbs measures

This chapter is devoted to establishing new comparison theorems for Gibbs measures that substantially extend the range of applicability of the classical Dobrushin comparison theorem, the main tool behind the proofs of the results presented in the previous two chapters for the analysis of filtering algorithms in high dimension. The novel toolbox will be used to extend the analysis of the block particle filter given in Chapter 4 to the case where spatial and temporal ergodicity are treated on a different footing. This chapter is based on the paper [41].

6.1 Motivations

The analysis of the block particle filter in Chapter 4 and the analysis of the localized Gibbs sampler particle filter in Chapter 5 rely heavily on the Dobrushin comparison theorem introduced in Section 2.4, which is a powerful tool to obtain dimension-free estimates on the difference between the marginals of Gibbs measures ρ and $\tilde{\rho}$ in terms of the single site conditional distributions $\rho(X^j \in dx^j | X^{I\setminus\{j\}} = x^{I\setminus\{j\}})$ and $\tilde{\rho}(X^j \in dx^j | X^{I\setminus\{j\}} = x^{I\setminus\{j\}})$.

In order to ensure decay of correlations, Theorem 4.2 and Theorem 5.4 (the main results of Chapter 4 and Chapter 5, respectively) impose a weak interactions assumption ($\varepsilon \leq p^v \leq \varepsilon^{-1}$ for $\varepsilon > \varepsilon_0$) that is dictated by the comparison theorem. As explained in Section 4.4.2, this assumption is unsatisfactory already at the *qualitative* level: it limits not only the spatial interactions (as is needed to ensure decay of correlations) but also the dynamics in time. Overcoming this unnatural restriction requires a generalized version of the comparison theorem, which is one of the main motivation for the results developed in this chapter.

More generally, aside from the filtering framework considered in the previous chapters, the Dobrushin comparison theorem has proved to be useful to establish numerous properties of Gibbs measures, including uniqueness, decay of correlations, global Markov properties, and analyticity [27, 45, 25], as well as functional inequalities and concentration of measure properties [29, 32, 67].

Despite this broad array of applications, the range of applicability of the Dobrushin comparison theorem proves to be somewhat limited. This can already be seen in the easiest qualitative consequence of this result: the comparison theorem implies uniqueness of the Gibbs measure under the well-known Dobrushin uniqueness criterion [18]. Unfortunately, this criterion is restrictive: even in models where uniqueness can be established by explicit computation, the Dobrushin uniqueness criterion holds only in a small subset of the natural parameter space (see, e.g., [64] for examples). This suggests that the Dobrushin comparison theorem is a rather blunt tool. On the other hand, it is also known that the Dobrushin uniqueness criterion can be substantially improved: this was accomplished in Dobrushin and Shlosman [17] by considering a local description in terms of larger blocks $\rho(X^J \in dx^J | X^{I\setminus J} = x^{I\setminus J})$ instead of the single site specification $\rho(X^j \in dx^j | X^{I\setminus\{j\}} = x^{I\setminus\{j\}})$. In this manner, it is possible in many cases to capture a large part of or even the entire uniqueness region. The uniqueness results of Dobrushin and Shlosman were further generalized by Weitz [64], who developed remarkably general combinatorial criteria for uniqueness. However, while the proofs of Dobrushin-Shlosman and Weitz also provide some information on decay of correlations, they do not provide an analogue of the powerful general-purpose machinery that the Dobrushin comparison theorem yields in its more restrictive setting.

The general aim of the present chapter is to fill this gap. Our main results (Theorem 6.4 and Theorem 6.12) provide a direct generalization of the Dobrushin comparison theorem to the much more general setting considered by Weitz [64], substantially extending the range of applicability of the classical comparison theorem.

While the original comparison theorem is an immediate consequence of our main result (Corollary 6.6), the classical proof that is based on the "method of estimates" does not appear to extend easily beyond the single site setting. We therefore develop a different, though certainly related, method of proof that systematically exploits the connection of Markov chains. In particular, our main results are derived from a more general comparison theorem for Markov chains that is applied to a suitably defined family of Gibbs samplers. The proofs of the new comparison theorems are contained in Appendix C, Sections C.1-C.5.

As an application of the generalized comparison theorems, in Section 6.4 we present an improved analysis of the block particle filter introduced in Chapter 4. The proof of this result is provided in Appendix C, Section C.6.

6.2 Setting and notation

We begin by introducing the basic setting that will be used throughout this section.

Sites and configurations

Let I be a finite or countably infinite set of *sites*. Each subset $J \subseteq I$ is called a *region*; the set of finite regions will be denoted as

$$\mathcal{I} := \{J \subseteq I : \operatorname{card} J < \infty\}.$$

To each site $i \in I$ is associated a measurable space $\mathbb{S}^i$, the *local state space*. A *configuration* is an assignment $x_i \in \mathbb{S}^i$ to each site $i \in I$. The set of all configurations $\mathbb{S}$, and the set $\mathbb{S}^J$ of configurations in a given region $J \subseteq I$, are defined as

$$\mathbb{S} := \prod_{i \in I} \mathbb{S}^i, \qquad \mathbb{S}^J := \prod_{i \in J} \mathbb{S}^i.$$

For $x = (x_i)_{i \in I} \in \mathbb{S}$, we denote by $x^J := (x_i)_{i \in J} \in \mathbb{S}^J$ the natural projection on $\mathbb{S}^J$. When $J \cap K = \varnothing$, we define $z = x^J y^K \in \mathbb{S}^{J \cup K}$ such that $z^J = x^J$ and $z^K = y^K$.

Local functions

A function $f : \mathbb{S} \to \bar{\mathbb{R}}$ is said to be *J-local* if $f(x) = f(z)$ whenever $x^J = z^J$, that is, if $f(x)$ depends on x^J only. The function f is said to be *local* if it is J-local for some finite region $J \in \mathcal{I}$. When I is a finite set, every function is local. When I is infinite, however, we will frequently restrict attention to local functions. More generally, we will consider a class of "nearly" local functions to be defined presently.

Given any function $f : \mathbb{S} \to \bar{\mathbb{R}}$, let us define for $J \in \mathcal{I}$ and $x \in \mathbb{S}$ the J-local function

$$f_x^J(z) := f(z^J x^{I \setminus J}).$$

Then f is called *quasilocal* if it can be approximated pointwise by the local functions f_x^J:

$$\lim_{J \in \mathcal{I}} |f_x^J(z) - f(z)| = 0 \quad \text{for all } x, z \in \mathbb{S},$$

where $\lim_{J \in \mathcal{I}} a_J$ denotes the limit of the net $(a_J)_{J \in \mathcal{I}}$ where $\mathcal{I}$ is directed by inclusion $\subseteq$ (equivalently, $a_J \to 0$ if and only if $a_{J_i} \to 0$ for every sequence $J_1, J_2, \ldots \in \mathcal{I}$ such that $J_1 \subseteq J_2 \subseteq \cdots$ and $\bigcup_i J_i = I$). Let us note that this notion is slightly weaker than the conventional notion of quasilocality used, for example, in [27].

Metrics

In the sequel, we fix for each $i \in I$ a metric η_i on $\mathbb{S}^i$ (we assume throughout that η_i is measurable as a function on $\mathbb{S}^i \times \mathbb{S}^i$). We will write $\|\eta_i\| = \sup_{x,z} \eta_i(x, z)$.

Given a function $f : \mathbb{S} \to \bar{\mathbb{R}}$ and $i \in I$, we define

$$\operatorname{osc}_i f := \sup_{x,z \in \mathbb{S} : x^{I \setminus \{i\}} = z^{I \setminus \{i\}}} \frac{|f(x) - f(z)|}{\eta_i(x_i, z_i)}.$$

The quantity $\operatorname{osc}_i f$ measures the variability of $f(x)$ with respect to the variable x_i.

Matrices

The calculus of possibly infinite nonnegative matrices will appear repeatedly in the sequel. Given matrices $A = (A_{ij})_{i,j \in I}$ and $B = (B_{ij})_{i,j \in I}$ with nonnegative entries $A_{ij} \geq 0$ and $B_{ij} \geq 0$, the matrix product is defined as usual by

$$(AB)_{ij} = \sum_{k \in I} A_{ik} B_{kj}.$$

This quantity is well defined as the terms in the sum are all nonnegative, but $(AB)_{ij}$ may possibly take the value $+\infty$. As long as we consider only nonnegative matrices, all the usual rules of matrix multiplication extend to infinite matrices provided that we allow entries with the value $+\infty$ and that we use the convention $+\infty \cdot 0 = 0$ (this follows from the Fubini-Tonelli theorem, cf. [20, Chapter 4]). In particular, the matrix powers A^k, $k \geq 1$ are well defined, and we define $A^0 = I$ where $I := (\mathbf{1}_{i=j})_{i,j \in I}$ denotes the identity matrix. We will write $A < \infty$ if the nonnegative matrix A satisfies $A_{ij} < \infty$ for every $i, j \in I$.

Kernels, covers, local structure

Recall that a *transition kernel* γ from a measurable space $(\Omega, \mathcal{F})$ to a measurable space $(\Omega', \mathcal{F}')$ is a map $\gamma : \Omega \times \mathcal{F}' \to \bar{\mathbb{R}}$ such that $\omega \mapsto \gamma_\omega(A)$ is a measurable function for each $A \in \mathcal{F}'$ and $\gamma_\omega(\cdot)$ is a probability measure for each $\omega \in \Omega$, cf. [31]. Given a probability measure μ on Ω and function f on Ω', we define as usual the probability measure $(\mu\gamma)(A) = \int \mu(d\omega)\gamma_\omega(A)$ on Ω' and function $(\gamma f)(\omega) = \int \gamma_\omega(d\omega')f(\omega')$ on Ω. A transition kernel γ between product spaces is called *quasilocal* if γf is quasilocal for every bounded and measurable quasilocal function f.

Our interest throughout this chapter is in models of random configurations, described by a probability measure μ on $\mathbb{S}$. We would like to understand the properties of such models based on their *local* structure. A natural way to express the local structure in a finite region $J \in \mathcal{I}$ is to consider the conditional distribution $\gamma_x^J(dz^J) = \mu(X^J \in dz^J | X^{I \setminus J} = x^{I \setminus J})$ of the configuration in J given a fixed configuration $x^{I \setminus J}$ for the sites outside J: conceptually, γ^J describes how the sites in J "interact" with the sites outside J. The conditional distribution γ^J is a transition kernel from $\mathbb{S}$ to $\mathbb{S}^J$. To obtain a complete local description of the model, we must consider a class of finite regions J that covers the entire set of sites I. Let us call a collection of regions $\mathcal{J} \subseteq \mathcal{I}$ a *cover* of I if every site $i \in I$ is contained in at least one element of $\mathcal{J}$ (note that, by definition, a cover contains only finite regions). Given any cover $\mathcal{J}$, the collection $(\gamma^J)_{J \in \mathcal{J}}$ provides a local description of the model.

In fact, our main results will hold in a somewhat more general setting than is described above. Let μ be a probability measure on $\mathbb{S}$ and γ^J be transition kernel from $\mathbb{S}$ to $\mathbb{S}^J$. We say that μ is γ^J*-invariant* if for every bounded measurable function f

$$\int \mu(dx)\, f(x) = \int \mu(dx)\, \gamma_x^J(dz^J)\, f(z^J x^{I \setminus J});$$

by a slight abuse of notation, we will also write $\mu f = \mu \gamma^J f^J$. This means that if the configuration x is drawn according to μ, then its distribution is left unchanged if we replace the configuration x^J inside the region J by a random sample from the distribution γ_x^J, keeping the configuration $x^{I \setminus J}$ outside J fixed. Our main results will be formulated in terms of a collection of transition kernels $(\gamma^J)_{J \in \mathcal{J}}$ such that $\mathcal{J}$ is a cover of I and such that μ is γ^J-invariant for every $J \in \mathcal{J}$. If we choose $\gamma_x^J(dz^J) = \mu(X^J \in dz^J | X^{I \setminus J} = x^{I \setminus J})$ as above, then the γ^J-invariance of μ holds by construction [31, Theorem 6.4]; however, any family of γ^J-invariant kernels will suffice for the validity of our main results.

Remark 6.1 (Gibbs measures and specifications)**.** *The idea that the collection* $(\gamma^J)_{J\in\mathcal{J}}$ *provides a natural description of high-dimensional probability distributions is prevalent in many applications. In fact, in statistical mechanics, the model is usually* defined *in terms of such a family. To this end, one fixes a priori a family of transition kernels* $(\gamma^J)_{J\in\mathcal{J}}$*, called a* specification*, that describes the local structure of the model. The definition of* γ^J *is done directly in terms of the parameters of the problem (the potentials that define the physical interactions, or the local constraints that define the combinatorial structure). A measure* μ *on* $\mathbb{S}$ *is called a* Gibbs measure *for the given specification if* $\mu(X^J \in dz^J | X^{I\setminus J} = x^{I\setminus J}) = \gamma_x^J(dz^J)$ *for every* $J \in \mathcal{J}$*. The existence of a Gibbs measure allows to define the model* μ *in terms of the specification. It may happen that there are multiple Gibbs measures for the same specification: the significance of this phenomenon is the presence of a* phase transition*, akin to the transition of water from liquid to solid at the freezing point. As the construction of Gibbs measures from specifications is not essential for the validity or applicability of our results, we omit further details. We refer to [27, 45, 64] for extensive discussion, examples, and references.*

6.3 General comparison theorem

Let ρ and $\tilde{\rho}$ be probability measures on the space of configurations $\mathbb{S}$. Our main result, Theorem 6.4 below, provides a powerful tool to obtain quantitative bounds on the difference between ρ and $\tilde{\rho}$ in terms of their local structure. Before we can state our results, we must first introduce some basic notions. Our terminology is inspired by Weitz [64].

As was explained above, the local description of a probability measure ρ on $\mathbb{S}$ will be provided in terms of a family of transition kernels. We formalize this as follows.

Definition 6.2. *A* local update rule *for* ρ *is a collection* $(\gamma^J)_{J\in\mathcal{J}}$ *where* $\mathcal{J}$ *is a cover of* I*,* γ^J *is a transition kernel from* $\mathbb{S}$ *to* $\mathbb{S}^J$ *and* ρ *is* γ^J*-invariant for every* $J \in \mathcal{J}$*.*

In order to compare two measures ρ and $\tilde{\rho}$ on the basis of their local update rules $(\gamma^J)_{J\in\mathcal{J}}$ and $(\tilde{\gamma}^J)_{J\in\mathcal{J}}$, we must quantify two separate effects. On the one hand, we must understand how the two models differ locally: that is, we must quantify how γ_x^J and $\tilde{\gamma}_x^J$ differ when acting on the same configuration x. On the other hand, we must understand how perturbations to the local update rule in different regions interact: to this end, we will quantify the extent to which γ_x^J and γ_z^J differ for different configurations x, z. Both effects will be addressed by introducing a suitable family of couplings. Recall that a probability measure Q on a product space $\Omega \times \Omega$ is called a *coupling* of probability measures μ, ν on Ω if its marginals coincide with μ, ν, that is, $Q(\,\cdot\, \times \Omega) = \mu$ and $Q(\Omega \times \,\cdot\,) = \nu$.

Definition 6.3. *A* coupled update rule *for* $(\rho, \tilde{\rho})$ *is a collection* $(\gamma^J, \tilde{\gamma}^J, Q^J, \hat{Q}^J)_{J\in\mathcal{J}}$*, where* $\mathcal{J}$ *is a cover of* I*, such that the following properties hold:*

1. $(\gamma^J)_{J\in\mathcal{J}}$ *and* $(\tilde{\gamma}^J)_{J\in\mathcal{J}}$ *are local update rules for* ρ *and* $\tilde{\rho}$*, respectively.*

2. $Q^J_{x,z}$ *is a coupling of* γ^J_x, γ^J_z *for every* $J \in \mathcal{J}$ *and* $x, z \in \mathbb{S}$ *with* $\mathrm{card}\{i : x_i \neq z_i\} = 1$.

3. $\hat{Q}^J_x$ *is a coupling of* $\gamma^J_x, \tilde{\gamma}^J_x$ *for every* $J \in \mathcal{J}$ *and* $x \in \mathbb{S}$.

We can now state our main result. The proof will be given in Appendix C, Sections C.1–C.3.

Theorem 6.4 (General comparison theorem, main result)**.** *Let* $\mathcal{J}$ *be a cover of* I*, let* $(w_J)_{J\in\mathcal{J}}$ *be a family of strictly positive weights, and let* $(\gamma^J, \tilde{\gamma}^J, Q^J, \hat{Q}^J)_{J\in\mathcal{J}}$ *be a coupled update rule for* $(\rho, \tilde{\rho})$*. Define for* $i, j \in I$

$$W_{ij} := \mathbf{1}_{i=j} \sum_{J\in\mathcal{J}: i\in J} w_J,$$

$$R_{ij} := \sup_{\substack{x,z\in\mathbb{S}:\\ x^{I\setminus\{j\}}=z^{I\setminus\{j\}}}} \frac{1}{\eta_j(x_j, z_j)} \sum_{J\in\mathcal{J}: i\in J} w_J\, Q^J_{x,z}\eta_i,$$

$$a_j := \sum_{J\in\mathcal{J}: j\in J} w_J \int^* \tilde{\rho}(dx)\, \hat{Q}^J_x \eta_j.$$

Assume that γ^J *is quasilocal for every* $J \in \mathcal{J}$*, and that*

$$W_{ii} \leq 1 \quad \textit{and} \quad \lim_{n\to\infty} \sum_{j\in I} (I - W + R)^n_{ij}\, (\rho \otimes \tilde{\rho})\eta_j = 0 \quad \textit{for all } i \in I. \tag{6.1}$$

Then we have

$$|\rho f - \tilde{\rho} f| \leq \sum_{i,j\in I} \mathrm{osc}_i f\, D_{ij}\, W^{-1}_{jj} a_j \quad \textit{where} \quad D := \sum_{n=0}^{\infty} (W^{-1}R)^n,$$

for any bounded and measurable quasilocal function f *such that* $\mathrm{osc}_i f < \infty$ *for all* $i \in I$.

Remark 6.5. *While it is essential in the proof that* γ^J *and* $\tilde{\gamma}^J$ *are transition kernels, we do not require that* Q^J *and* $\hat{Q}^J$ *are transition kernels in Definition 6.3, that is, the couplings* $Q^J_{x,z}$ *and* $\hat{Q}^J_x$ *need not be measurable as functions of* x, z*. It is for this reason that the coefficients* a_j *are defined in terms of an outer integral rather than an ordinary integral [53]:*

$$\int^* f(x)\, \rho(dx) := \inf \left\{ \int g(x)\, \rho(dx) : f \leq g,\ g \textit{ is measurable} \right\}.$$

When $x \mapsto \hat{Q}^J_x \eta_j$ *is measurable this issue can be disregarded. In practice measurability will hold in all but pathological cases, but may not always be trivial to prove. We therefore allow for nonmeasurable couplings for sake of technical convenience, so that it is not necessary to check measurability of the coupled updates when applying Theorem 6.4.*

We will presently formulate a number of special cases and extensions of Theorem 6.4 that may be useful in different settings. A detailed application is presented in Section 6.4, where we improve the analysis of the block particle filter given in Chapter 4.

6.3.1 The classical comparison theorem

The original comparison theorem of Dobrushin [18, Theorem 3] and its commonly used formulation due to Föllmer [24] (i.e., Theorem 2.11) correspond to the special case of Theorem 6.4 where the cover $\mathcal{J} = \mathcal{J}_{\rm s} := \{\{i\} : i \in I\}$ consists of single sites. For example, the main result of [24] follows readily from Theorem 6.4 under a mild regularity assumption. To formulate it, recall that the *Wasserstein distance* $d_\eta(\mu,\nu)$ between probability measures μ and ν on a measurable space Ω with respect to a measurable metric η is defined as

$$d_\eta(\mu,\nu) := \inf_{\substack{Q(\cdot\times\Omega)=\mu\\ Q(\Omega\times\cdot)=\nu}} Q\eta,$$

where the infimum is taken over probability measures Q on $\Omega\times\Omega$ with the given marginals μ and ν. We now obtain the following classical result (cf. [24] and [25, Remark 2.17]).

Corollary 6.6 ([24]). *Assume $\mathbb{S}^i$ is Polish and η_i is lower-semicontinuous for all $i \in I$. Let $(\gamma^{\{i\}})_{i\in I}$ and $(\tilde\gamma^{\{i\}})_{i\in I}$ be local update rules for ρ and $\tilde\rho$, respectively, and let*

$$C_{ij} := \sup_{\substack{x,z\in\mathbb{S}:\\ x^{I\setminus\{j\}}=z^{I\setminus\{j\}}}} \frac{d_{\eta_i}(\gamma_x^{\{i\}},\gamma_z^{\{i\}})}{\eta_j(x_j,z_j)}, \qquad b_j := \int^* \tilde\rho(dx)\, d_{\eta_j}(\gamma_x^{\{j\}},\tilde\gamma_x^{\{j\}}).$$

Assume that $\gamma^{\{i\}}$ is quasilocal for every $i \in I$, and that

$$\lim_{n\to\infty} \sum_{j\in I} C_{ij}^n (\rho\otimes\tilde\rho)\eta_j = 0 \quad \textit{for all } i \in I.$$

Then we have

$$|\rho f - \tilde\rho f| \le \sum_{i,j\in I} \mathrm{osc}_i f\, D_{ij}\, b_j \quad \textit{where} \quad D := \sum_{n=0}^\infty C^n,$$

for any bounded and measurable quasilocal function f such that $\mathrm{osc}_i f < \infty$ for all $i \in I$.

If $Q_{x,z}^{\{i\}}$ and $\hat Q_x^{\{i\}}$ are minimizers in the definition of $d_{\eta_i}(\gamma_x^{\{i\}},\gamma_z^{\{i\}})$ and $d_{\eta_i}(\gamma_x^{\{i\}},\tilde\gamma_x^{\{i\}})$, respectively, and if we let $\mathcal{J} = \mathcal{J}_{\rm s}$ and $w_{\{i\}} = 1$ for all $i \in I$, then Corollary 6.6 follows immediately from Theorem 6.4. For simplicity, we have imposed the mild topological regularity assumption on $\mathbb{S}^i$ and η_i to ensure the existence of minimizers [62, Theorem 4.1] (when minimizers do not exist, it is possible with some more work to obtain a similar result by using near-optimal couplings in Theorem 6.4). Let us note that when $\eta_i(x,z) = \mathbf{1}_{x\neq z}$ is the trivial metric, the Wasserstein distance reduces to the total variation distance

$$d_\eta(\mu,\nu) = \frac{1}{2}\|\mu-\nu\| := \frac{1}{2}\sup_{f:\|f\|\le 1} |\mu f - \nu f| \quad \text{when} \quad \eta(x,z) = \mathbf{1}_{x\neq z},$$

and an optimal coupling exists in any measurable space [18, p. 472]. Thus in this case no regularity assumptions are needed, and Corollary 6.6 reduces to the textbook version of the comparison theorem that appears, e.g., in [27, Theorem 8.20] or [45, Theorem V.2.2].

While the classical comparison theorem of Corollary 6.6 follows from our main result, it should be emphasized that the single site assumption $\mathcal{J} = \mathcal{J}_{\mathrm{s}}$ is a significant restriction. The general statement of Theorem 6.4 constitutes a crucial improvement that substantially extends the range of applicability of the comparison method, as the application to the block particle filter demonstrates. Let us also note that the proofs in [18, 24], based on the "method of estimates," do not appear to extend easily beyond the single site setting. We use a different (though related) method of proof that systematically exploits the connection with Markov chains (Appendix C).

6.3.2 Alternative assumptions

The key assumption of Theorem 6.4 is (6.1). The aim of the present section is to obtain a number of useful alternatives to assumption (6.1) that are easily verified in practice.

We begin by defining the notion of a tempered measure [25, Remark 2.17].

Definition 6.7. *A probability measure μ on $\mathbb{S}$ is called* $x^\star$-tempered *if*

$$\sup_{i\in I} \int \mu(dx)\, \eta_i(x_i, x_i^\star) < \infty.$$

In the sequel $x^\star \in \mathbb{S}$ will be considered fixed and μ will be called tempered.

It is often the case in practice that the collection of metrics is uniformly bounded, that is, $\sup_i \|\eta_i\| < \infty$. In this case, every probability measure on $\mathbb{S}$ is trivially tempered. However, the restriction to tempered measures may be essential when the spaces $\mathbb{S}^i$ are noncompact (see, for example, [18, section 5] for a simple but illuminating example).

Let us recall that a norm $\|\cdot\|$ defined on an algebra of square (possibly infinite) matrices is called a *matrix norm* if $\|AB\| \le \|A\|\,\|B\|$. We also recall that the matrix norms $\|\cdot\|_\infty$ and $\|\cdot\|_1$ are defined for nonnegative matrices $A = (A_{ij})_{i,j\in I}$ as

$$\|A\|_\infty := \sup_{i\in I} \sum_{j\in I} A_{ij}, \qquad\qquad \|A\|_1 := \sup_{j\in I} \sum_{i\in I} A_{ij}.$$

The following result collects various useful alternatives to (6.1). It is proved in Section C.4 in Appendix C.

Corollary 6.8 (Alternatives to assumption (6.1))**.** *Suppose that ρ and $\tilde\rho$ are tempered. Then the conclusion of Theorem 6.4 remains valid when the assumption (6.1) is replaced by one of the following:*

1. $\operatorname{card} I < \infty$ *and* $D < \infty$.

2. $\operatorname{card} I < \infty$, $R < \infty$, *and* $\|(W^{-1}R)^n\| < 1$ *for some matrix norm* $\|\cdot\|$ *and* $n \geq 1$.

3. $\sup_i W_{ii} < \infty$ *and* $\|W^{-1}R\|_\infty < 1$.

4. $\sup_i W_{ii} < \infty$, $\|RW^{-1}\|_\infty < \infty$, *and* $\|(RW^{-1})^n\|_\infty < 1$ *for some* $n \geq 1$.

5. $\sup_i W_{ii} < \infty$, $\sum_i \|\eta_i\| < \infty$, *and* $\|RW^{-1}\|_1 < 1$.

6. $\sup_i W_{ii} < \infty$, *there exists a metric* m *on* I *such that* $\sup\{m(i,j) : R_{ij} > 0\} < \infty$ *and* $\sup_i \sum_j e^{-\beta m(i,j)} < \infty$ *for all* $\beta > 0$, *and* $\|RW^{-1}\|_1 < 1$.

The conditions of Corollary 6.8 are closely related to the uniqueness problem for Gibbs measures. Suppose that the collection of quasilocal transition kernels $(\gamma^J)_{J\in\mathcal{J}}$ is a local update rule for ρ. It is natural to ask whether ρ is the *unique* measure that admits $(\gamma^J)_{J\in\mathcal{J}}$ as a local update rule (see the remark at the end of Section 6.2). We now observe that uniqueness is a necessary condition for the conclusion of Theorem 6.4. Indeed, let $\tilde\rho$ be another measure that admits the same local update rule. If (6.1) holds, we can apply Theorem 6.4 with $\tilde\gamma^J = \gamma^J$ and $a_j = 0$ to conclude that $\tilde\rho = \rho$. In particular, $\sum_j (I - W + R)^n_{ij} \to 0$ in Theorem 6.4 evidently implies uniqueness in the class of tempered measures.

Of course, the point of Theorem 6.4 is that it provides a quantitative tool that goes far beyond qualitative uniqueness questions. It is therefore interesting to note that this single result nonetheless captures many of the uniqueness conditions that are used in the literature. In Corollary 6.8, Condition 3 is precisely the "influence on a site" condition of Weitz [64, Theorem 2.5] (our setting is even more general in that we do not require bounded-range interactions as is essential in [64]). Conditions 5 and 6 constitute a slight strengthening (see below) of the "influence of a site" condition of Weitz [64, Theorem 2.7] under summable metric or subexponential graph assumptions, in the spirit of the classical uniqueness condition of Dobrushin and Shlosman [17]. In the finite setting with single site updates, Condition 2 is in the spirit of [22] and Condition 4 is in the spirit of [21].

On the other hand, we can now see that Theorem 6.4 provides a crucial improvement over the classical comparison theorem. The single site setting of Corollary 6.6 corresponds essentially to the original Dobrushin uniqueness regime [18]. It is well known that this setting is restrictive, in that it captures only a small part of the parameter space where uniqueness of Gibbs measures holds. It is precisely for this reason that Dobrushin and Shlosman introduced their improved uniqueness criterion in terms of larger blocks [17], which in many cases allows to capture a large part of or even the entire uniqueness region; see [64, section 5] for examples. The generalized comparison Theorem 6.4 in terms of larger blocks can therefore be fruitfully applied to a much larger and more natural class of models than the classical comparison theorem. This point is further emphasized in the context of the application to the block particle filter in Section 6.4.

Remark 6.9. *The "influence of a site" condition* $\|RW^{-1}\|_1 < 1$ *that appears in Corollary 6.8 is slightly stronger than the corresponding condition of Dobrushin-*

Shlosman [17] and Weitz [64, Theorem 2.7]. Writing out the definition of R, we find that our condition reads

$$\|RW^{-1}\|_1 = \sup_{j\in I} W_{jj}^{-1} \sum_{i\in I} \sup_{\substack{x,z\in\mathbb{S}:\\ x^{I\setminus\{j\}}=z^{I\setminus\{j\}}}} \frac{1}{\eta_j(x_j,z_j)} \sum_{J\in\mathfrak{J}:i\in J} w_J\, Q^J_{x,z}\eta_i < 1,$$

while the condition of [64, Theorem 2.7] (which extends the condition of [17]) reads

$$\sup_{j\in I} W_{jj}^{-1} \sup_{\substack{x,z\in\mathbb{S}:\\ x^{I\setminus\{j\}}=z^{I\setminus\{j\}}}} \frac{1}{\eta_j(x_j,z_j)} \sum_{i\in I} \sum_{J\in\mathfrak{J}:i\in J} w_J\, Q^J_{x,z}\eta_i < 1.$$

The latter is slightly weaker as the sum over sites i appears inside the supremum over configurations x, z. While the distinction between these conditions is inessential in many applications, there do exist situations in which the weaker condition yields an essential improvement, see, e.g., [64, section 5.3]. In such problems, Theorem 6.4 is not only limited by the stronger uniqueness condition but could also lead to poor quantitative bounds, as the comparison bound is itself expressed in terms of the uniform influence coefficients R_{ij}.

It could therefore be of interest to develop comparison theorems that are able to exploit the finer structure that is present in the weaker uniqueness condition. In fact, the proof of Theorem 6.4 already indicates a natural approach to such improved bounds. However, the resulting comparison theorems are necessarily nonlinear *in that the action of the matrix R is replaced by a nonlinear operator* R*. The nonlinear expressions are somewhat difficult to handle in practice, and as we do not at present have a compelling application for such bounds we do not pursue this direction here. However, for completeness, we will briefly sketch at the end of Section C.2 how such bounds can be obtained.*

6.3.3 A one-sided comparison theorem

As was discussed in Section 6.2, it is natural in many applications to describe high-dimensional probability distributions in terms of local conditional probabilities of the form $\mu(X^J \in dz^J | X^{I\setminus J} = x^{I\setminus J})$. This is in essence a static picture, where we describe the behavior of each local region J given that the configuration of the remaining sites $I\setminus J$ is frozen. In models that possess dynamics, this description is not very natural. In this setting, each site $i \in I$ occurs at a given time $\tau(i)$, and its state is only determined by the configuration of sites $j \in I$ in the past and present $\tau(j) \le \tau(i)$, but not by the future. For example, the model might be defined as a high-dimensional Markov chain whose description is naturally given in terms of one-sided conditional probabilities (see, e.g., [23]). It is therefore interesting to note that the original comparison theorem of Dobrushin [18] is actually more general than Corollary 6.6 in that it is applicable both in the static and dynamic settings (see the one-sided Dobrushin comparison theorem, Theorem 2.12). We presently develop an analogous generalization to Theorem 6.4.

For the purposes of this section, we assume that we are given a function $\tau : I \to \mathbb{Z}$ that assigns to each site $i \in I$ an integer index $\tau(i)$. We define

$$I_{\leq k} := \{i \in I : \tau(i) \leq k\}, \qquad \mathbb{S}_{\leq k} := \mathbb{S}^{I_{\leq k}},$$

and for any probability measure ρ on $\mathbb{S}$ we denote by $\rho_{\leq k}$ the marginal distribution on $\mathbb{S}_{\leq k}$.

Definition 6.10. *A* one-sided local update rule *for ρ is a collection $(\gamma^J)_{J\in\mathcal{J}}$ where*

1. *$\mathcal{J}$ is a cover of I such that $\min_{i\in J} \tau(i) = \max_{i\in J} \tau(i) =: \tau(J)$ for every $J \in \mathcal{J}$.*
2. *γ^J is a transition kernel from $\mathbb{S}_{\leq\tau(J)}$ to $\mathbb{S}^J$.*
3. *$\rho_{\leq\tau(J)}$ is γ^J-invariant for every $J \in \mathcal{J}$.*

The canonical example of a one-sided local update rule is to consider the one-sided conditional distributions $\gamma^J_x(dz^J) = \rho(X^J \in dz^J | X^{I_{\leq\tau(J)}\setminus J} = x^{I_{\leq\tau(J)}\setminus J})$. This situation is particularly useful in the investigation of interacting Markov chains, cf. [18, 23], where $\tau(j)$ denotes the time index of the site j and we condition only on the past and present, but not on the future.

Definition 6.11. *A* one-sided coupled update rule *for $(\rho, \tilde\rho)$ is a collection of transition kernels $(\gamma^J, \tilde\gamma^J, Q^J, \hat Q^J)_{J\in\mathcal{J}}$ such that the following hold:*

1. *$(\gamma^J)_{J\in\mathcal{J}}$ and $(\tilde\gamma^J)_{J\in\mathcal{J}}$ are one-sided local update rules for ρ and $\tilde\rho$, respectively.*
2. *$Q^J_{x,z}$ is a coupling of γ^J_x, γ^J_z for $J \in \mathcal{J}$ and $x, z \in \mathbb{S}_{\leq\tau(J)}$ with $\mathrm{card}\{i : x_i \neq z_i\} = 1$.*
3. *$\hat Q^J_x$ is a coupling of $\gamma^J_x, \tilde\gamma^J_x$ for $J \in \mathcal{J}$ and $x \in \mathbb{S}_{\leq\tau(J)}$.*

We can now state a one-sided counterpart to Theorem 6.4, which will be proved in Section C.5.

Theorem 6.12 (General comparison theorem, one-sided)**.** *Let $(\gamma^J, \tilde\gamma^J, Q^J, \hat Q^J)_{J\in\mathcal{J}}$ be a one-sided coupled update rule for $(\rho, \tilde\rho)$, and let $(w_J)_{J\in\mathcal{J}}$ be a family of strictly positive weights. Define the matrices W and R and the vector a as in Theorem 6.4. Assume that γ^J is quasilocal for every $J \in \mathcal{J}$, that*

$$\sum_{j\in I} D_{ij}\,(\rho\otimes\tilde\rho)\eta_j < \infty \quad \text{for all } i \in I \qquad \text{where} \qquad D := \sum_{n=0}^{\infty} (W^{-1}R)^n, \tag{6.2}$$

and that (6.1) holds. Then we have

$$|\rho f - \tilde\rho f| \leq \sum_{i,j\in I} \mathrm{osc}_i f\, D_{ij}\, W^{-1}_{jj} a_j$$

for any bounded and measurable quasilocal function f such that $\mathrm{osc}_i f < \infty$ for all $i \in I$.

Let us remark that the result of Theorem 6.12 is formally the same as that of Theorem 6.4, except that we have changed the nature of the update rules used in the definition of the coefficients. We also require a further assumption (6.2) in addition to assumption (6.1) of Theorem 6.4, but this is not restrictive in practice: in particular, it is readily verified that the conclusion of Theorem 6.12 also holds under any of the conditions of Corollary 6.8.

6.4 Application: block particle filter

Our original motivation for developing the generalized comparison theorems of this chapter was the investigation of algorithms for filtering in high dimension. In this section we state a result that improve *qualitatively* Theorem 4.2—the main result of Chapter 4—on the analysis of the block particle filter.

We assume to be in the same set up of Chapter 4, and we refer to Section 4.4.2 therein for a discussion that motivates the importance of the following theorem. The proof of this result, which relies crucially on the generalized comparison theorems developed in this chapter, is provided in Appendix C, Section C.6.

Theorem 6.13 (Block particle filter, improved version of Theorem 4.2)**.** *For any $0 < \delta < 1$ there exists $0 < \varepsilon_0 < 1$, depending only on δ and Δ, such that the following holds. Suppose there exist $\varepsilon_0 < \varepsilon < 1$ and $0 < \kappa < 1$ so that*

$$\begin{aligned} \varepsilon q^v(x^v, z^v) \le p^v(x, z^v) \le \varepsilon^{-1} q^v(x^v, z^v), \\ \delta \le q^v(x^v, z^v) \le \delta^{-1}, \\ \kappa \le g^v(x^v, y^v) \le \kappa^{-1} \end{aligned}$$

for every $v \in V$, $x, z \in \mathbb{X}$, $y \in \mathbb{Y}$, where $q^v : \mathbb{X}^v \times \mathbb{X}^v \to \bar{\mathbb{R}}_+$ is a transition density with respect to ψ^v. Then for every $n \ge 0$, $\sigma \in \mathbb{X}$, $K \in \mathcal{K}$ and $J \subseteq K$ we have

$$|||\pi_n^\sigma - \hat{\pi}_n^\sigma|||_J \le \alpha \operatorname{card} J \left[e^{-\beta_1 d(J, \partial K)} + \frac{e^{\beta_2 |\mathcal{K}|_\infty}}{N^\gamma} \right],$$

where $0 < \gamma \le \frac{1}{2}$ and $0 < \alpha, \beta_1, \beta_2 < \infty$ depend only on δ, ε, κ, r, Δ, and $\Delta_{\mathcal{K}}$.

In Theorem 6.13, the parameter ε controls the spatial correlations while the parameter δ controls the temporal correlations (in contrast to Theorem 4.2, where both are controlled simultaneously by ε). The key point is that δ can be arbitrary, and only ε must lie above the threshold ε_0. That the threshold ε_0 depends on δ is natural: the more ergodic the dynamics, the more spatial interactions can be tolerated without losing decay of correlations.

The proof of Theorem 4.2 was based on repeated application of the classical Dobrushin comparison theorem (Corollary 6.6). While there are some significant differences between the details of the proofs, the essential improvement that makes it possible to prove Theorem 6.13 is that we can now exploit the generalized comparison theorem (Theorem 6.4), which enables us to treat the spatial and temporal degrees of freedom on a different footing (see Section C.6).

Chapter 7

Nonlinear filtering in infinite dimension

This chapter is devoted to showing that filtering in infinite dimension is *qualitatively* different from filtering in finite dimension. We show that new phenomena arise in the infinite-dimensional setting, specifically that inheritance of ergodicity (in the form of stability or decay of correlations) can undergo a phase transition in the signal-to-noise ratio. The qualitative setting of this chapter is complementary to the quantitative framework previously considered in this thesis. The material here presented is taken from the paper [42], which further develops this set of ideas by providing conditions to guarantee inheritance of ergodicity.

7.1 Motivations

In Chapter 4 and Chapter 5 we have shown that local filtering algorithms can attain dimension-free approximation errors in high-dimensional models that exhibit conditional decay of correlations. The natural tool to capture and exploit decay of correlations is given by the Dobrushin comparison theorem, and in Chapter 6 we extended this machinery by introducing more general comparison theorems.

The framework developed in the previous chapters is complementary in nature to the one developed in the present chapter: the former provide *quantitative* estimates under strong ('high-temperature') assumptions, while the latter focuses on the *qualitative* understanding of ergodic properties of the filter distribution. In fact, as discussed in Section 4.4.1, the local analysis of filtering algorithms that we have developed relies on the crucial assumption that we can establish proper forms of filter stability and decay of correlations. Presently, we address the fundamental question of the inheritance of such properties upon conditioning.

To discuss the topic of this chapter, let $(X_k, Y_k)_{k\geq 0}$ be a bivariate Markov chain of the kind considered in this thesis. 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

$$\pi_k := \mathbf{P}(X_k \in \cdot\,|Y_1, \ldots, Y_k).$$

Under standard assumptions on the observation structure, 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 $k \to \infty$, 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 1957 [5], and a proof was provided by Kunita in 1971 [33]. 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 7.1 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 [57, 52] 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* [54, 56, 58, 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

[1] Surprisingly, the counterexample (intended for a different purpose) appears in Blackwell's own paper [5].

models in data assimilation problems that arise in areas such as weather forecasting or geophysics [49]. The aim of this chapter 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.

To model a filtering problem in infinite dimension we extend the framework introduced in Section 4.1. We now 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 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 $V = \mathbb{Z}^d$.

In Section 7.2 we review the classical results on the inheritance of filter stability, and we discuss Blackwell's Example 7.1. In Section 7.3 we introduce the canonical infinite-dimensional model that will be studied in this chapter, and in Section 7.4 we investigate the natural infinite-dimensional version of Blackwell's Example. 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, in Section 7.5 we conjecture that this might be the case under a symmetry breaking assumption. We refer to the paper [42] for a more detailed discussion on this conjecture, and for some positive results that go towards proving it.

In Section 7.6 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 [66, 26]. Here the natural question of interest is whether the spatial mixing properties of random fields are inherited by conditioning on local observations. Again, we refer to the paper [42] for more details on the matter.

7.2 Inheritance of ergodicity: classical results

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 that slightly generalizes the one introduced in Chapter 3, and we introduce and discuss the basic ergodicity question to be studied.

Throughout this chapter, we model dynamics with partial information as a hidden Markov models where $(X_k, Y_k)_{k\geq 0}$ is a Markov chain 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 $(X_k)_{k\geq 0}$. More general settings could also be considered, see [51] for instance.

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 7.3 below). The nonlinear filter is defined as the regular conditional probability

$$\pi_k := \mathbf{P}(X_k \in \cdot\, | Y_1, \ldots, Y_k).$$

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 7.3 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 chapter.

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 $k < 0$.

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

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

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

$$|\mathbf{P}(X_k \in A|X_0, Y_1, \ldots, Y_k) - \mathbf{P}(X_k \in A|Y_1, \ldots, Y_k)| \xrightarrow{k\to\infty} 0 \quad \text{in } L^1$$

for every measurable set A: that is, the conditional distribution of X_k given the observed data "forgets" the initial condition X_0 as $k \to \infty$. 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 [5].

Example 7.1 (Blackwell's example)**.** *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 $k \geq 0$

$$\begin{aligned}\mathbf{P}(X_k = 1|X_0, Y_1, \ldots, Y_k) &= \mathbf{1}_{X_k=1},\\ \mathbf{P}(X_k = 1|Y_1, \ldots, Y_k) &= 1/2.\end{aligned}$$

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:

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

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 7.1 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 7.1. Unfortunately, this is still not the case in complete generality, as is demonstrated by an esoteric counterexample in [59]. However, the conclusion does hold if we use a stronger *uniform* notion of stability.

Theorem 7.2 (Inheritance of stability [57])**.** *Suppose that the following hold.*

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

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

2. *The observations are* nondegenerate *in the sense*

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

Then the filter is uniformly stable *in the sense*

$$\sup_A |\mathbf{P}(X_k \in A|X_0, Y_1, \ldots, Y_k) - \mathbf{P}(X_k \in A|Y_1, \ldots, Y_k)| \xrightarrow{k\to\infty} 0 \quad \textit{in } L^1.$$

This result, together with the mathematical theory behind its proof 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 chapter, we will see that new phenomena arise in the infinite-dimensional setting.

Remark 7.3 (Different formulations of filter stability)**.** *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 chapter 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 $\mathbf{P}^\mu$ 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

$$\mathbf{P}^\mu(X_k \in \cdot) \xrightarrow{k\to\infty} \lambda \quad \textit{for every } \mu$$

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

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

for a suitable class of measures μ, ν, γ and functions f. The formulation that we consider in this chapter 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 [57, 52].

A different and perhaps more natural formulation dates back to Blackwell [5] and Kunita [33]. 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. [59, 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 chapter, cf. [33, 48, 9, 13, 59].

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 chapter, a fact that is exploited heavily in the theory of [57, 52].

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 7.4 (On observability)**.** *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 [54, 56, 58] in terms of nonlinear notions of* observability*. Such results are in some sense the opposite of Theorem 7.2: 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.*

7.3 The infinite-dimensional model

The aim of this chapter is to show that new phenomena arise in filtering theory 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 7.2, 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 (this model represents a generalization of the finite-dimensional model considered in Section 4.1).

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,

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

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

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

where

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

Such a model should be viewed as a hidden Markov model counterpart of probabilistic cellular automata [35] or interacting particle systems [36] 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 $(X_k^v, Y_k^v)_{k\ge0}$ 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 7.2 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 $g(e,\cdot) \ne g(e',\cdot)$, 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 (cf. Proposition 2.14). This immediately rules out the possibility that the observations are nondegenerate in the sense of Theorem 7.2. It is precisely this problem that lies at the heart of the

[2] The 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. [49].

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); see Section 2.4. 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 7.5 (Observations: the problem in infinite dimension)**.** *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 7.2 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 $k < \infty$, which rules out total variation convergence (see [52, Example 2.3] for a simple illustration). However, this issue is surmounted in [52] using a form of localization: by performing the analysis of Theorem 7.2 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 7.2 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 7.6 (On infinite-dimensional models)**.** *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., [52]). See also Section 2.4.*

At the same time, it should be noted that even in finite-dimensional systems where results such as Theorem 7.2 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 Chapter 4 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.

7.4 A conditional phase transition

We now develop a simple example of the general infinite-dimensional setting of Section 7.3 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 7.1 above).

Throughout this section,

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

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

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

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 7.3. 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 $p \le 1/2$, which we will do from now on.

The model that we have constructed is evidently a direct extension of Example 7.1 to infinite dimension. As in Example 7.1, 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 7.2 is satisfied. When $p = 0$, it is easily seen by the same reasoning as in Example 7.1 that the filter is not stable. However, in Example 7.1 the addition of observation noise with error probability $p > 0$ would yield nondegenerate observations, and thus filter stability by Theorem 7.2. 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 7.7 (Inheritance of stability, phase transition)**.** *For the model of this section, there exist constants $0 < p_\star \le p^\star < 1/2$ such that the filter is stable for $p^\star < p \le 1/2$ and is not stable for $0 \le p < p_\star$.*

We refer to Appendix D for the proof of Theorem 7.7. The proof relies on standard tools from statistical mechanics [7, 27]: a Peierls argument for the low noise regime and a Dobrushin contraction method for the high noise regime.

Remark 7.8. *We naturally believe that one can choose $p_\star = p^\star$ in Theorem 7.7, but we did not succeed in proving that. The proof yields some explicit bounds on $p_\star$ and $p^\star$.*

Theorem 7.7 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.

7.5 Conjecture on inheritance of stability

Theorem 7.7 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, as seen in Chapter 4 and Chapter 5.

It is not difficult to understand the mechanism that causes the filter to be unstable in Theorem 7.7. 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 7.1. In the finite-dimensional case, however, Theorem 7.2 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 7.7 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 7.9. *Let $(X_k, Y_k)_{k\in\mathbb{Z}}$ be a stationary infinite-dimensional hidden Markov model as in Section 7.3 with $X_k \in \{-1,1\}^{\mathbb{Z}}$ and with $Y_k \in \{-1,1\}^{\mathbb{Z}}$ of the form*

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

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 $Y \perp\!\!\!\perp X$; we will therefore assume $p \neq \frac{1}{2}$ in the sequel). Thus any mechanism of the type exhibited by Theorem 7.7 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. This phenomenon was made precise in the papers [54, 56, 58]. While the theory developed in these papers is closely related to the symmetry breaking properties that we aim to exploit here, these results are not satisfactory in infinite dimension.

In the paper [42] we extend such observability arguments to *translation-invariant* systems in infinite dimension by exploiting a technique from multidimensional ergodic theory [12]. Somewhat surprisingly, the problem proves to be more tractable in the continuous-time setting, for which will establish validity of the natural analogue of Conjecture 7.9. In its original discrete time formulation, however, our ultimate result falls short of establishing Conjecture 7.9 even for translation-invariant models. Nonetheless, the theory developed here provides one possible mechanism for symmetry breaking in conditional ergodic theory.

7.6 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, for the theory developed in Chapter 4 and Chapter 5 it is of significant interest 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 [66, 26].

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 Section 7.4 and in Chapter 4 and Chapter 5 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 chapter is organized as follows. In Section 7.6.1, we recall some basic notions from the theory of Markov random fields. In Section 7.6.2, we develop basic properties of conditional random fields and introduce some of the relevant questions.

7.6.1 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 $v \in \mathbb{Z}^d$ and that each X_v takes values in a finite set E.

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

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

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 7.10. *$X = (X_v)_{v\in\mathbb{Z}^d}$ 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* [27, Chapter 1] (cf. Remark 6.1).

Definition 7.11. *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 measurable $A \subseteq E^{\mathbb{Z}^d}$ 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 $\mathscr{G}(\gamma)$.*

Example 7.12. *Standard constructions of Markov random fields arise in statistical mechanics in the following manner. Let $\psi_v : E \to \mathbb{R}$ and $\varphi_{\{v,w\}} : E \times E \to \mathbb{R}$ 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 can be 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 $\mathscr{G}(\gamma)$ 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 $\mathscr{G}(\gamma)$ is closely related to the spatial mixing properties of the associated random fields, as is shown by the following result [27, 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 $\mathscr{G}(\gamma)$, then $\lambda \mathbf{P} + (1 - \lambda)\mathbf{Q}$ is also in $\mathscr{G}(\gamma)$ for $0 \le \lambda \le 1$ [27, Chapter 7]; thus $\mathscr{G}(\gamma)$ is a convex set, and a random field is called *extremal* if it is an extreme point of this set.

Theorem 7.13. *For a given specification γ, the following hold.*

1. *Existence of a random field: $\mathscr{G}(\gamma) \neq \varnothing$.*
2. *Uniqueness $\Leftrightarrow$ uniform mixing: $|\mathscr{G}(\gamma)| = 1$ iff a random field in $\mathscr{G}(\gamma)$ satisfies*[3,4]

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

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

[3] Here 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 $\mathscr{G}(\gamma)$, but the equivalence between uniqueness and uniform mixing is false if a null set is omitted in the supremum over x.

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

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 7.13 is a direct spatial analogue of the stability property of a Markov chain introduced in Section 7.2. 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).

7.6.2 Conjecture on inheritance of decay of correlations

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 7.3, 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*.

Remark 7.14. *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 7.17). 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 7.15. *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.

Proposition 7.16. *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* $\mathscr{G}(\gamma^Y)$ *a.s.*
3. (X, Y) *is conditionally mixing iff* $\mathbf{P}(X \in \cdot\,|Y)$ *is extremal in* $\mathscr{G}(\gamma^Y)$ *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 $\mathscr{G}(\gamma^Y)$ 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. 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 $\mathscr{G}(\gamma^Y)$ 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

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

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

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

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

$$\lim_{n\to\infty} \mathbf{E}^y|\mathbf{P}^y(X \in A|X_{W_n^c}) - \mathbf{P}^y(X \in A)| = \mathbf{E}^y|\mathbf{P}^y(X \in A|\textstyle\bigcap_n \sigma\{X_{W_n^c}\}) - \mathbf{P}^y(X \in A)|.$$

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 $\mathbf{P}(X \in \cdot\,|Y)$. □

Proposition 7.16 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 $\mathscr{G}(\gamma)$, when is $\mathbf{P}(X \in \cdot\,|Y)$ extremal in $\mathscr{G}(\gamma^Y)$ a.s.?
2. If $|\mathscr{G}(\gamma)| = 1$, when is $|\mathscr{G}(\gamma^Y)| = 1$ 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 7.13 that $|\mathscr{G}(\gamma^Y)| = 1$ a.s. implies the conditional mixing property. The stronger conclusion $|\mathscr{G}(\gamma^Y)| = 1$ 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 [26].

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 7.4.

Example 7.17 (Inheritance of decay of correlations, phase transition)**.** *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 7.4. 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 7.16 holds in this setting (see also Remark 7.14 above and Remark 7.18 below).

We can now proceed identically as in the proof of Theorem 7.7 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 7.7 in the first place: the model $(X^v_k, Y^v_k)_{k,v\in\mathbb{Z}}$ 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 7.18. *The use of edge observations in Example 7.17 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 $\tilde Y_v = \big(X_v X_{v+(0,1)} \xi_{\{v,v+(0,1)\}}, X_v X_{v+(1,0)} \xi_{\{v,v+(1,0)\}}\big)$, where X_v and $\xi_{\{v,w\}}$ are as in Example 7.17. 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. [26, p. 6]). For example, we can formulate the natural analogue of Conjecture 7.9.

Conjecture 7.19. *Let $(X_v, Y_v)_{v\in\mathbb{Z}^2}$ be a hidden Markov field with $E = F = \{-1, 1\}$ and*

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

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, in [42] we 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.

Appendix A

Block particle filter: proofs

The goal of this appendix is to prove Theorem 4.2. We refer to Section 4.5 for an overview of the main ideas in the proof that we are going to present.

Theorem 4.2 yields a bound on $|||\pi_n^\mu - \hat{\pi}_n^\mu|||_J$. As

$$|||\pi_n^\mu - \hat{\pi}_n^\mu|||_J \leq \underbrace{|||\pi_n^\mu - \tilde{\pi}_n^\mu|||_J}_{\text{bias}} + \underbrace{|||\tilde{\pi}_n^\mu - \hat{\pi}_n^\mu|||_J}_{\text{variance}}$$

it suffices to bound each term in this inequality. As was explained in Section 4.5.1, the first term quantifies the bias of the block particle filter, while the second term quantifies the variance of the random sampling. The bias term will be bounded in Theorem A.12 below, while the variance will be bounded in Theorem A.21. The combination of these two results immediately yields Theorem 4.2.

The Dobrushin comparison method, as discussed in Section 4.5.2, is the main workhorse of our proof. To use this method, we must be able to bound the quantities C_{ij}, b_j, and D_{ij} that appear in the Dobrushin comparison theorem (Theorem 2.11). We have already introduced in Section 2.3 and Section 2.4 some elementary lemmas for this purpose. We also need the following lemma to bounds C_{ij}.

Lemma A.1 (Minorization condition)**.** *Let $\nu, \nu', \gamma, \gamma'$ be probability measures on a measurable space $(E, \mathcal{E})$, and let $\varepsilon > 0$ be such that $\nu(A) \geq \varepsilon\gamma(A)$ and $\nu'(A) \geq \varepsilon\gamma'(A)$ for every measurable set A. Then*

$$|||\nu - \nu'||| \leq 2(1-\varepsilon) + \varepsilon|||\gamma - \gamma'|||.$$

In particular, if $\gamma = \gamma'$, then $|||\nu - \nu'||| \leq 2(1-\varepsilon)$. The same conclusion holds if the $|||\cdot|||$-norm is replaced by the $\|\cdot\|$-norm.

Proof. As $\mu = (1-\varepsilon)^{-1}(\nu - \varepsilon\gamma)$ and $\mu' = (1-\varepsilon)^{-1}(\nu' - \varepsilon\gamma')$ are probability measures and $\nu - \nu' = (1-\varepsilon)(\mu - \mu') + \varepsilon(\gamma - \gamma')$, the result follows readily. □

A.1 Local stability of the filter

The main goal of this section is to prove a local stability bound for the nonlinear filter. We begin, however, by introducing a number of objects that will appear several times in the sequel.

For any probability measure μ on $\mathbb{X}$ and $x, z \in \mathbb{X}$, $v \in V$, we define

$$\begin{aligned}\mu^v_{x,z}(A) &:= \mathbf{P}^\mu(X_0^v \in A | X_0^{V\setminus\{v\}} = x^{V\setminus\{v\}},\ X_1 = z) \\ &= \frac{\int \mathbf{1}_A(x^v) \prod_{w\in N(v)} p^w(x, z^w)\, \mu^v_x(dx^v)}{\int \prod_{w\in N(v)} p^w(x, z^w)\, \mu^v_x(dx^v)}\end{aligned}$$

(recall the notation $\mu^v_x := \mathbf{P}^\mu(X_0^v \in \cdot\, | X_0^{V\setminus\{v\}} = x^{V\setminus\{v\}})$ in Section 4.5.2). Let

$$C^\mu_{vv'} := \frac{1}{2} \sup_{z\in\mathbb{X}} \sup_{x,\tilde{x}\in\mathbb{X}: x^{V\setminus\{v'\}} = \tilde{x}^{V\setminus\{v'\}}} \|\mu^v_{x,z} - \mu^v_{\tilde{x},z}\|$$

for $v, v' \in V$. The quantity

$$\mathrm{Corr}(\mu, \beta) := \max_{v\in V} \sum_{v'\in V} e^{\beta d(v,v')} C^\mu_{vv'}$$

could be viewed as a measure of the degree of correlation decay of the measure μ at rate $\beta > 0$. It will turn out that this (not entirely obvious) measure of decay of correlations is precisely tuned to the needs of the proof of Theorem 4.2. This is due to the fact that the measures $\mu^v_{x,z}$ arise naturally when applying the Dobrushin comparison method to the smoothing distributions as discussed in Section 4.5.2.

Proposition A.2 (Local filter stability)**.** *Suppose there exists $\varepsilon > 0$ such that*

$$\varepsilon \le p^v(x, z^v) \le \varepsilon^{-1} \qquad \text{for all } v \in V,\ x, z \in \mathbb{X}.$$

Let μ, ν be probability measures on $\mathbb{X}$, and suppose that

$$\mathrm{Corr}(\mu, \beta) + 3(1 - \varepsilon^{2\Delta}) e^{2\beta r} \Delta^2 \le \frac{1}{2}$$

for a sufficiently small constant $\beta > 0$. Then we have

$$\begin{aligned}&\|\mathsf{F}_n \cdots \mathsf{F}_{s+1}\mu - \mathsf{F}_n \cdots \mathsf{F}_{s+1}\nu\|_J \\ &\qquad \le 2e^{-\beta(n-s)} \sum_{v\in J} \max_{v'\in V} e^{-\beta d(v,v')} \sup_{x,z\in\mathbb{X}} \|\mu^{v'}_{x,z} - \nu^{v'}_{x,z}\|\end{aligned}$$

for every $J \subseteq V$ and $s < n$.

Remark A.3. *There is nothing magical about the constant $1/2$ in the decay of correlations assumption; any constant $c < 1$ would work at the expense of a constant $1/(1-c)$ rather than 2 in the filter stability bound. As our methods are not expected to yield tight quantitative bounds, we have taken the liberty to fix various constants of this sort throughout the following sections for aesthetic purposes.*

Remark A.4. *Note that by Lemma 2.9*

$$\|\mu^{v'}_{x,z} - \nu^{v'}_{x,z}\| \le \frac{2}{\varepsilon^{2\Delta}} \|\mu^{v'}_x - \nu^{v'}_x\|.$$

This yields a slightly cleaner bound in Proposition A.2 with a worse constant. For our purposes, however, it will be just as easy to bound $\|\mu^{v'}_{x,z} - \nu^{v'}_{x,z}\|$ directly.

Proof. Define the smoothing distributions

$$\rho = \mathbf{P}^\mu(X_0, \dots, X_n \in \cdot\,|Y_1, \dots, Y_n),$$
$$\tilde\rho = \mathbf{P}^\nu(X_0, \dots, X_n \in \cdot\,|Y_1, \dots, Y_n).$$

We will apply Theorem 2.11 to $\rho, \tilde\rho$ with $I = \{0, \dots, n\} \times V$ and $\mathbb{S} = \mathbb{X}^{n+1}$ as discussed in 4.5.2. To this end, we must bound the quantities C_{ij} and b_j. We begin by bounding C_{ij} with $i = (k, v)$ and $j = (k', v')$. We distinguish three cases.

Case $k = 0$. The key observation in this case is that $\rho^i_x = \mu^v_{x_0,x_1}$ by the Markov property (or by direct computation). Note that as $\operatorname{card} N(v) \le \Delta$, we have

$$\mu^v_{x,z}(A) = \frac{\int \mathbf{1}_A(x^v) \prod_{w\in N(v)} p^w(x, z^w)\, \mu^v_x(dx^v)}{\int \prod_{w\in N(v)} p^w(x, z^w)\, \mu^v_x(dx^v)} \ge \varepsilon^{2\Delta}\, \mu^v_x(A),$$

so $\|\mu^v_{x,z} - \mu^v_{x,z'}\| \le 2(1 - \varepsilon^{2\Delta})$ for any $z, z' \in \mathbb{X}$ by Lemma A.1. Therefore

$$C_{ij} \le \begin{cases} C^\mu_{vv'} & \text{if } k' = 0, \\ 1 - \varepsilon^{2\Delta} & \text{if } k' = 1 \text{ and } v' \in N(v), \\ 0 & \text{otherwise.} \end{cases}$$

This evidently implies that

$$\sum_{(k',v')\in I} e^{\beta k'} e^{\beta d(v,v')} C_{(0,v)(k',v')} \le \operatorname{Corr}(\mu, \beta) + (1 - \varepsilon^{2\Delta}) e^{\beta(r+1)} \Delta.$$

Case $0 < k < n$. Now we have (cf. Section 4.5.2)

$$\rho^i_x(A) = \frac{\int \mathbf{1}_A(x^v_k)\, p^v(x_{k-1}, x^v_k)\, g^v(x^v_k, Y^v_k) \prod_{w\in N(v)} p^w(x_k, x^w_{k+1})\, \psi^v(dx^v_k)}{\int p^v(x_{k-1}, x^v_k)\, g^v(x^v_k, Y^v_k) \prod_{w\in N(v)} p^w(x_k, x^w_{k+1})\, \psi^v(dx^v_k)}.$$

By inspection, ρ^i_x does not depend on $x^{v'}_{k'}$ except in the following cases: $k' = k - 1$ and $v' \in N(v)$; $k' = k + 1$ and $v' \in N(v)$; $k' = k$ and $v' \in \bigcup_{w\in N(v)} N(w)$. As

$$\rho^i_x(A) \ge \varepsilon^{2\Delta} \frac{\int \mathbf{1}_A(x^v_k)\, p^v(x_{k-1}, x^v_k)\, g^v(x^v_k, Y^v_k)\, \psi^v(dx^v_k)}{\int p^v(x_{k-1}, x^v_k)\, g^v(x^v_k, Y^v_k)\, \psi^v(dx^v_k)}$$

as well as

$$\rho^i_x(A) \ge \varepsilon^2 \frac{\int \mathbf{1}_A(x^v_k)\, g^v(x^v_k, Y^v_k) \prod_{w\in N(v)} p^w(x_k, x^w_{k+1})\, \psi^v(dx^v_k)}{\int g^v(x^v_k, Y^v_k) \prod_{w\in N(v)} p^w(x_k, x^w_{k+1})\, \psi^v(dx^v_k)},$$

we can use Lemma A.1 to estimate

$$C_{ij} \le \begin{cases} 1-\varepsilon^2 & \text{if } k' = k-1 \text{ and } v' \in N(v), \\ 1-\varepsilon^{2\Delta} & \text{if } k' = k+1 \text{ and } v' \in N(v), \\ 1-\varepsilon^{2\Delta} & \text{if } k' = k \text{ and } v' \in \bigcup_{w \in N(v)} N(w), \\ 0 & \text{otherwise.} \end{cases}$$

This yields

$$\begin{aligned} \sum_{(k',v') \in I} e^{\beta|k-k'|} e^{\beta d(v,v')} C_{(k,v)(k',v')} &\le (1-\varepsilon^{2\Delta})\{e^{2\beta r}\Delta^2 + 2e^{\beta(r+1)}\Delta\} \\ &\le 3(1-\varepsilon^{2\Delta}) e^{2\beta r}\Delta^2, \end{aligned}$$

where we have used that $r \ge 1$ and $\Delta \ge 1$ in the last inequality.

Case $k = n$. Now we have

$$\begin{aligned} \rho_x^i(A) &= \frac{\int \mathbf{1}_A(x_n^v)\, p^v(x_{n-1}, x_n^v)\, g^v(x_n^v, Y_n^v)\, \psi^v(dx_n^v)}{\int p^v(x_{n-1}, x_n^v)\, g^v(x_n^v, Y_n^v)\, \psi^v(dx_n^v)} \\ &\ge \varepsilon^2 \frac{\int \mathbf{1}_A(x_n^v)\, g^v(x_n^v, Y_n^v)\, \psi^v(dx_n^v)}{\int g^v(x_n^v, Y_n^v)\, \psi^v(dx_n^v)}, \end{aligned}$$

and we obtain precisely as above

$$C_{ij} \le \begin{cases} 1-\varepsilon^2 & \text{if } k' = n-1 \text{ and } v' \in N(v), \\ 0 & \text{otherwise.} \end{cases}$$

We therefore find

$$\sum_{(k',v') \in I} e^{\beta|k-k'|} e^{\beta d(v,v')} C_{(n,v)(k',v')} \le (1-\varepsilon^2) e^{\beta(r+1)}\Delta.$$

Combining the above three cases and the assumption of the Proposition yields

$$\max_{(k,v) \in I} \sum_{(k',v') \in I} e^{\beta\{|k-k'|+d(v,v')\}} C_{(k,v)(k',v')} \le \frac{1}{2}.$$

Thus Lemma 2.13 gives

$$\max_{(k,v) \in I} \sum_{(k',v') \in I} e^{\beta\{|k-k'|+d(v,v')\}} D_{(k,v)(k',v')} \le 2.$$

Now consider the quantities b_j in Theorem 2.11. By the Markov property, it is evident that $\rho_x^i = \tilde{\rho}_x^i$ whenever $i = (k, v)$ with $k \ge 1$. On the other hand, for $k = 0$ we obtain $\rho_x^i = \mu_{x_0,x_1}^v$ and $\tilde{\rho}_x^i = \nu_{x_0,x_1}^v$. Applying Theorem 2.11 therefore yields

$$\|\pi_n^\mu - \pi_n^\nu\|_J = \|\rho - \tilde{\rho}\|_{\{n\} \times J} \le \sum_{v \in J} \sum_{v' \in V} D_{(n,v)(0,v')} \sup_{x,z \in \mathbb{X}} \|\mu_{x,z}^{v'} - \nu_{x,z}^{v'}\|.$$

However, note that

$$\begin{aligned}
&\sum_{v'\in V} D_{(n,v)(0,v')} \sup_{x,z\in\mathbb{X}} \|\mu^{v'}_{x,z} - \nu^{v'}_{x,z}\| \\
&\quad = e^{-\beta n} \sum_{v'\in V} e^{\beta\{n+d(v,v')\}} D_{(n,v)(0,v')}\, e^{-\beta d(v,v')} \sup_{x,z\in\mathbb{X}} \|\mu^{v'}_{x,z} - \nu^{v'}_{x,z}\| \\
&\quad \le 2e^{-\beta n} \max_{v'\in V} e^{-\beta d(v,v')} \sup_{x,z\in\mathbb{X}} \|\mu^{v'}_{x,z} - \nu^{v'}_{x,z}\|,
\end{aligned}$$

using the above estimate on the matrix D. Substituting this into the bound for $\|\pi^\mu_n - \pi^\nu_n\|_J$ yields the statement of the Proposition for the special case $s = 0$.

To obtain the result for any $s < n$, note that $\mathsf{F}_n \cdots \mathsf{F}_{s+1}\mu$ and π^μ_{n-s} differ only in that a different sequence of observations ($Y_{s+1}, \ldots, Y_n$ versus $Y_1, \ldots, Y_{n-s}$) is used in the computation of these quantities. As our bound holds uniformly in the observation sequence, however, the general result follows immediately. □

As a corollary of Proposition A.2, let us derive a simple filter stability statement that illustrates the role of decay of correlations (this will not be used elsewhere).

Corollary A.5 (Filter stability)**.** *Suppose there exists $\varepsilon > 0$ such that*

$$\varepsilon \le p^v(x, z^v) \le \varepsilon^{-1} \qquad \text{for all } v \in V,\ x, z \in \mathbb{X},$$

and such that

$$\varepsilon > \varepsilon_0 = \left(1 - \frac{1}{6\Delta^2}\right)^{1/2\Delta}.$$

Then for any probability measures μ, ν on $\mathbb{X}$ and $J \subseteq V$, $n \ge 0$, we have

$$\|\pi^\mu_n - \pi^\nu_n\|_J \le 4 \operatorname{card} J\, \gamma^{n/2r},$$

where $\gamma = 6\Delta^2(1 - \varepsilon^{2\Delta}) < 1$.

Proof. We first apply Proposition A.2 with $\mu = \delta_x$. Then $\mathrm{Corr}(\mu, \beta) = 0$ for any $\beta > 0$. Choosing $\beta = -(2r)^{-1} \log \gamma > 0$, we find that

$$\mathrm{Corr}(\mu, \beta) + 3(1 - \varepsilon^{2\Delta}) e^{2\beta r} \Delta^2 = \frac{1}{2},$$

so that the assumption of Proposition A.2 is satisfied. Therefore,

$$\|\pi^x_n - \pi^\nu_n\|_J \le 4 \operatorname{card} J\, e^{-\beta n} = 4 \operatorname{card} J\, \gamma^{n/2r}.$$

To obtain the result for arbitrary μ, note that

$$\begin{aligned}
\pi^\mu_n(A) &= \mathbf{P}^\mu(X_n \in A | Y_1, \ldots, Y_n) \\
&= \mathbf{E}^\mu(\mathbf{P}^\mu(X_n \in A | X_0, Y_1, \ldots, Y_n) | Y_1, \ldots, Y_n) \\
&= \mathbf{E}^\mu(\pi^{\delta_{X_0}}_n(A) | Y_1, \ldots, Y_n).
\end{aligned}$$

Therefore, by Jensen's inequality,

$$\|\pi^\mu_n - \pi^\nu_n\|_J \le \mathbf{E}^\mu(\|\pi^{\delta_{X_0}}_n - \pi^\nu_n\|_J | Y_1, \ldots, Y_n) \le \sup_{x\in\mathbb{X}} \|\pi^x_n - \pi^\nu_n\|_J,$$

which yields the result. □

While Proposition A.2 requires a decay of correlations assumption on the initial condition ($\mathrm{Corr}(\mu, \beta)$ must be sufficiently small), Corollary A.5 works for any initial condition provided that $\varepsilon > \varepsilon_0$ is sufficiently large (which is necessary in general, see Section 4.4.1). Thus no assumption is needed on the initial condition if we want to show only that the filter is stable in time. On the other hand, Proposition A.2 controls not only the stability in time, but also the spatial accumulation of error between μ and ν by virtue of the damping factor $e^{-\beta d(v,v')}$: the decay of correlations property of the initial condition is essential to obtain this type of local control. The latter is of central importance if we wish to obtain local error bounds for filter approximations that are uniform in time and in the model dimension.

A.2 The block projection error

The proof of a time-uniform error bound between π_n^μ and $\tilde{\pi}_n^\mu$ requires two ingredients: we need the filter stability property of π_n^μ, developed in the previous section, in order to mitigate the accumulation of approximation errors over time; and we need to control the approximation error between π_n^μ and $\tilde{\pi}_n^\mu$ in one time step. The latter is the purpose of this section.

We will in fact consider two separate cases. To control the total error $\|\pi_n^\mu - \tilde{\pi}_n^\mu\|_J$, we need to consider the one-step error made in each time step $s = 1, \ldots, n$. For time steps $s < n$ (for which the error is dissipated by the stability of the filter), the error must be measured in terms of the quantities that appear in Proposition A.2: that is, we must control $\|(\mathsf{F}_s\nu)^v_{x,z} - (\tilde{\mathsf{F}}_s\nu)^v_{x,z}\|$. On the other hand, in the last time step $s = n$, we must control directly $\|\mathsf{F}_n\nu - \tilde{\mathsf{F}}_n\nu\|_J$. While the proofs of these cases are quite similar, each much be considered separately in the following.

We begin by bounding the error in time steps $s < n$.

Proposition A.6 (Block error, $s < n$)**.** *Suppose there exists $\varepsilon > 0$ such that*

$$\varepsilon \le p^v(x, z^v) \le \varepsilon^{-1} \qquad \text{for all } v \in V,\ x, z \in \mathbb{X}.$$

Let ν be a probability measure on $\mathbb{X}$, and suppose that

$$\mathrm{Corr}(\nu, \beta) + (1 - \varepsilon^2)e^{\beta(r+1)}\Delta \le \frac{1}{2}$$

for a sufficiently small constant $\beta > 0$. Then we have

$$\sup_{x,z\in\mathbb{X}} \|(\mathsf{F}_s\nu)^v_{x,z} - (\tilde{\mathsf{F}}_s\nu)^v_{x,z}\| \le 4e^{-\beta}(1 - \varepsilon^{2\Delta})\, e^{-\beta d(v,\partial K)}$$

for every $s \in \mathbb{N}$, $K \in \mathcal{K}$ and $v \in K$.

This result makes precise the idea that was heuristically expressed in Section 4.3: if the measure ν possesses the decay of correlations property, then the error at site v incurred by applying the block filter rather than the true filter decays exponentially in the distance between v and the boundary of the block that it is in.

Proof. We begin by writing out the definitions

$$(\mathsf{F}_s\nu)(A) = \frac{\int \mathbf{1}_A(x) \prod_{w\in V} p^w(x_0,x^w)\, g^w(x^w,Y_s^w)\, \nu(dx_0)\, \psi(dx)}{\int \prod_{w\in V} p^w(x_0,x^w)\, g^w(x^w,Y_s^w)\, \nu(dx_0)\, \psi(dx)},$$
$$(\tilde{\mathsf{F}}_s\nu)(A) = \frac{\int \mathbf{1}_A(x) \prod_{K'\in\mathcal{K}} \big[\int \prod_{w\in K'} p^w(x_0,x^w)\, g^w(x^w,Y_s^w)\, \nu(dx_0)\big]\, \psi(dx)}{\int \prod_{K'\in\mathcal{K}} \big[\int \prod_{w\in K'} p^w(x_0,x^w)\, g^w(x^w,Y_s^w)\, \nu(dx_0)\big]\, \psi(dx)}.$$

Let us fix $K\in\mathcal{K}$, $v\in K$ throughout the proof. Then

$$(\mathsf{F}_s\nu)^v_x(A) = \frac{\int \mathbf{1}_A(x^v)\, g^v(x^v,Y_s^v) \prod_{w\in V} p^w(x_0,x^w)\, \nu(dx_0)\, \psi^v(dx^v)}{\int g^v(x^v,Y_s^v) \prod_{w\in V} p^w(x_0,x^w)\, \nu(dx_0)\, \psi^v(dx^v)},$$
$$(\tilde{\mathsf{F}}_s\nu)^v_x(A) = \frac{\int \mathbf{1}_A(x^v)\, g^v(x^v,Y_s^v) \prod_{w\in K} p^w(x_0,x^w)\, \nu(dx_0)\, \psi^v(dx^v)}{\int g^v(x^v,Y_s^v) \prod_{w\in K} p^w(x_0,x^w)\, \nu(dx_0)\, \psi^v(dx^v)}.$$

Define $I = (\{0\}\times V)\cup(1,v)$ and $\mathbb{S} = \mathbb{X}\times\mathbb{X}^v$, and the probability measures on $\mathbb{S}$

$$\rho(A) = \frac{\int \mathbf{1}_A(x_0,x^v)\, g^v(x^v,Y_s^v) \prod_{w\in V} p^w(x_0,x^w) \prod_{u\in N(v)} p^u(x,z^u)\, \nu(dx_0)\, \psi^v(dx^v)}{\int g^v(x^v,Y_s^v) \prod_{w\in V} p^w(x_0,x^w) \prod_{u\in N(v)} p^u(x,z^u)\, \nu(dx_0)\, \psi^v(dx^v)},$$
$$\tilde\rho(A) = \frac{\int \mathbf{1}_A(x_0,x^v)\, g^v(x^v,Y_s^v) \prod_{w\in K} p^w(x_0,x^w) \prod_{u\in N(v)} p^u(x,z^u)\, \nu(dx_0)\, \psi^v(dx^v)}{\int g^v(x^v,Y_s^v) \prod_{w\in K} p^w(x_0,x^w) \prod_{u\in N(v)} p^u(x,z^u)\, \nu(dx_0)\, \psi^v(dx^v)}.$$

Then we have by construction

$$\|(\mathsf{F}_s\nu)^v_{x,z} - (\tilde{\mathsf{F}}_s\nu)^v_{x,z}\| = \|\rho-\tilde\rho\|_{(1,v)}.$$

We will apply Theorem 2.11 to bound $\|\rho-\tilde\rho\|_{(1,v)}$. To this end, we must bound C_{ij} and b_i with $i=(k',v')$ and $j=(k'',v'')$. We distinguish two cases.

Case $k'=0$. In this case we have

$$\rho^i_{(x_0,x^v)}(A) = \frac{\int \mathbf{1}_A(x_0^{v'}) \prod_{w\in N(v')} p^w(x_0,x^w)\, \nu^{v'}_{x_0}(dx_0^{v'})}{\int \prod_{w\in N(v')} p^w(x_0,x^w)\, \nu^{v'}_{x_0}(dx_0^{v'})},$$
$$\tilde\rho^i_{(x_0,x^v)}(A) = \frac{\int \mathbf{1}_A(x_0^{v'}) \prod_{w\in N(v')\cap K} p^w(x_0,x^w)\, \nu^{v'}_{x_0}(dx_0^{v'})}{\int \prod_{w\in N(v')\cap K} p^w(x_0,x^w)\, \nu^{v'}_{x_0}(dx_0^{v'})}.$$

In particular, $\rho^i_{(x_0,x^v)} = \nu^{v'}_{x_0,x}$, so $C_{ij}\le C^\nu_{v'v''}$ if $k''=0$. Moreover, as

$$\rho^i_{(x_0,x^v)}(A) \ge \varepsilon^2 \frac{\int \mathbf{1}_A(x_0^{v'}) \prod_{w\in N(v')\setminus\{v\}} p^w(x_0,x^w)\, \nu^{v'}_{x_0}(dx_0^{v'})}{\int \prod_{w\in N(v')\setminus\{v\}} p^w(x_0,x^w)\, \nu^{v'}_{x_0}(dx_0^{v'})},$$

we have $C_{ij}\le 1-\varepsilon^2$ if $k''=1$ (so $v''=v$) and $v\in N(v')$ by Lemma A.1, and $C_{ij}=0$ otherwise. We therefore immediately obtain the estimate

$$\sum_{(k'',v'')\in I} e^{\beta k''} e^{\beta d(v',v'')} C_{(0,v')(k'',v'')} \le \mathrm{Corr}(\nu,\beta) + (1-\varepsilon^2)e^{\beta(r+1)}.$$

On the other hand, note that $\rho^i_{(x_0,x^v)} = \tilde{\rho}^i_{(x_0,x^v)}$ if $N(v') \subseteq K$, and that we have $\rho^i_{(x_0,x^v)} \geq \varepsilon^{2\Delta}\nu^{v'}_{x_0}$ and $\tilde{\rho}^i_{(x_0,x^v)} \geq \varepsilon^{2\Delta}\nu^{v'}_{x_0}$. Therefore, by Lemma A.1

$$b_i = \sup_{(x_0,x^v)\in\mathbb{S}} \|\rho^i_{(x_0,x^v)} - \tilde{\rho}^i_{(x_0,x^v)}\| \leq \begin{cases} 0 & \text{for } v' \in K\backslash\partial K, \\ 2(1-\varepsilon^{2\Delta}) & \text{otherwise.} \end{cases}$$

Case $k' = 1$. In this case we have

$$\rho^i_{(x_0,x^v)}(A) = \tilde{\rho}^i_{(x_0,x^v)}(A) = \frac{\int \mathbf{1}_A(x^v)\, g^v(x^v, Y^v_s)\, p^v(x_0,x^v) \prod_{u\in N(v)} p^u(x,z^u)\, \psi^v(dx^v)}{\int g^v(x^v, Y^v_s)\, p^v(x_0,x^v) \prod_{u\in N(v)} p^u(x,z^u)\, \psi^v(dx^v)}.$$

Thus $b_i = 0$, and estimating as above we obtain $C_{ij} \leq 1-\varepsilon^2$ whenever $k'' = 0$ and $v'' \in N(v)$, and $C_{ij} = 0$ otherwise. In particular, we obtain

$$\sum_{(k'',v'')\in I} e^{\beta|1-k''|} e^{\beta d(v,v'')} C_{(1,v)(k'',v'')} \leq (1-\varepsilon^2) e^{\beta(r+1)}\Delta.$$

Combining the above two cases and the assumption of the Proposition yields

$$\max_{(k',v')\in I} \sum_{(k'',v'')\in I} e^{\beta\{|k'-k''|+d(v',v'')\}} C_{(k',v')(k'',v'')} \leq \frac{1}{2}.$$

Applying Theorem 2.11 and Lemma 2.13 gives

$$\begin{aligned} \|(\mathsf{F}_s\nu)^v_{x,z} - (\tilde{\mathsf{F}}_s\nu)^v_{x,z}\| &= \|\rho - \tilde{\rho}\|_{(1,v)} \\ &\leq 2(1-\varepsilon^{2\Delta}) \sum_{v'\in V\backslash(K\backslash\partial K)} D_{(1,v)(0,v')} \\ &\leq 4e^{-\beta}(1-\varepsilon^{2\Delta}) e^{-\beta d(v,\partial K)}. \end{aligned}$$

As the choice of $x, z \in \mathbb{X}$ was arbitrary, the proof is complete. □

We now use a similar argument to bound the error in time step n.

Proposition A.7 (Block error, $s = n$)**.** *Suppose there exists $\varepsilon > 0$ such that*

$$\varepsilon \leq p^v(x,z^v) \leq \varepsilon^{-1} \qquad \text{for all } v \in V,\ x, z \in \mathbb{X}.$$

Let ν be a probability measure on $\mathbb{X}$, and suppose that

$$\mathrm{Corr}(\nu,\beta) + (1-\varepsilon^2)e^{\beta(r+1)}\Delta \leq \frac{1}{2}$$

for a sufficiently small constant $\beta > 0$. Then we have

$$\|\mathsf{F}_n\nu - \tilde{\mathsf{F}}_n\nu\|_J \leq 4e^{-\beta}(1-\varepsilon^{2\Delta})\, e^{-\beta d(J,\partial K)}\, \mathrm{card}\, J$$

for every $K \in \mathcal{K}$ and $J \subseteq K$.

Proof. Define $I = \{0,1\} \times V$ and $\mathbb{S} = \mathbb{X}^2$. Fix $K \in \mathcal{K}$, and let

$$\rho(A) = \frac{\int \mathbf{1}_A(x_0,x_1) \prod_{v\in V} p^v(x_0,x_1^v)\, g^v(x_1^v, Y_n^v)\, \nu(dx_0)\, \psi(dx_1)}{\int \prod_{v\in V} p^v(x_0,x_1^v)\, g^v(x_1^v, Y_n^v)\, \nu(dx_0)\, \psi(dx_1)},$$
$$\tilde\rho(A) = \frac{\int \mathbf{1}_A(x_0,x_1) \prod_{v\in K} p^v(x_0,x_1^v) \prod_{w\in V} g^w(x_1^w, Y_n^w)\, \nu(dx_0)\, \psi(dx_1)}{\int \prod_{v\in K} p^v(x_0,x_1^v) \prod_{w\in V} g^w(x_1^w, Y_n^w)\, \nu(dx_0)\, \psi(dx_1)}.$$

Then for any $J \subseteq K$, we have

$$\|\mathsf{F}_n\nu - \tilde{\mathsf{F}}_n\nu\|_J = \|\rho - \tilde\rho\|_{\{1\}\times J}.$$

We will apply Theorem 2.11 to bound $\|\rho - \tilde\rho\|_{\{1\}\times J}$. To this end, we must bound C_{ij} and b_i with $i = (k,v)$ and $j = (k',v')$. We distinguish two cases.

Case $k = 0$. In this case we have

$$\rho^i_x(A) = \frac{\int \mathbf{1}_A(x_0^v) \prod_{w\in N(v)} p^w(x_0,x_1^w)\, \nu^v_{x_0}(dx_0^v)}{\int \prod_{w\in N(v)} p^w(x_0,x_1^w)\, \nu^v_{x_0}(dx_0^v)},$$
$$\tilde\rho^i_x(A) = \frac{\int \mathbf{1}_A(x_0^v) \prod_{w\in N(v)\cap K} p^w(x_0,x_1^w)\, \nu^v_{x_0}(dx_0^v)}{\int \prod_{w\in N(v)\cap K} p^w(x_0,x_1^w)\, \nu^v_{x_0}(dx_0^v)}.$$

In particular, $\rho^i_x = \nu^v_{x_0,x_1}$, so $C_{ij} \le C^\nu_{vv'}$ if $k' = 0$. Moreover, as

$$\rho^i_x(A) \ge \varepsilon^2 \frac{\int \mathbf{1}_A(x_0^v) \prod_{w\in N(v)\setminus\{v'\}} p^w(x_0,x_1^w)\, \nu^v_{x_0}(dx_0^v)}{\int \prod_{w\in N(v)\setminus\{v'\}} p^w(x_0,x_1^w)\, \nu^v_{x_0}(dx_0^v)},$$

we have $C_{ij} \le 1-\varepsilon^2$ if $k' = 1$ and $v' \in N(v)$ by Lemma A.1, and $C_{ij} = 0$ otherwise. We therefore immediately obtain the estimate

$$\sum_{(k',v')\in I} e^{\beta k'} e^{\beta d(v,v')} C_{(0,v)(k',v')} \le \mathrm{Corr}(\nu,\beta) + (1-\varepsilon^2) e^{\beta(r+1)}\Delta.$$

On the other hand, note that $\rho^i_x = \tilde\rho^i_x$ if $N(v) \subseteq K$, and that we have $\rho^i_x \ge \varepsilon^{2\Delta}\nu^v_{x_0}$ and $\tilde\rho^i_x \ge \varepsilon^{2\Delta}\nu^v_{x_0}$. Therefore, we obtain by Lemma A.1

$$b_i = \sup_{x\in\mathbb{S}} \|\rho^i_x - \tilde\rho^i_x\| \le \begin{cases} 0 & \text{for } v \in K\setminus\partial K, \\ 2(1-\varepsilon^{2\Delta}) & \text{otherwise.} \end{cases}$$

Case $k = 1$. In this case we have

$$\rho^i_x(A) = \frac{\int \mathbf{1}_A(x_1^v)\, p^v(x_0,x_1^v)\, g^v(x_1^v, Y_n^v)\, \psi^v(dx_1^v)}{\int p^v(x_0,x_1^v)\, g^v(x_1^v, Y_n^v)\, \psi^v(dx_1^v)},$$

while $\tilde\rho^x_i = \rho^x_i$ if $v \in K$ and

$$\tilde\rho^i_x(A) = \frac{\int \mathbf{1}_A(x_1^v)\, g^v(x_1^v, Y_n^v)\, \psi^v(dx_1^v)}{\int g^v(x_1^v, Y_n^v)\, \psi^v(dx_1^v)},$$

otherwise. Thus we obtain from Lemma A.1

$$b_i = \sup_{x\in\mathbb{S}} \|\rho_x^i - \tilde{\rho}_x^i\| \le \begin{cases} 0 & \text{for } v \in K, \\ 2(1-\varepsilon^2) & \text{otherwise.} \end{cases}$$

On the other hand, we can readily estimate as above

$$\sum_{(k',v')\in I} e^{\beta|1-k'|} e^{\beta d(v,v')} C_{(1,v)(k',v')} \le (1-\varepsilon^2) e^{\beta(r+1)} \Delta.$$

Combining the above two cases and the assumption of the Proposition yields

$$\max_{(k,v)\in I} \sum_{(k',v')\in I} e^{\beta\{|k-k'|+d(v,v')\}} C_{(k,v)(k',v')} \le \frac{1}{2}.$$

Applying Theorem 2.11 and Lemma 2.13 gives

$$\begin{aligned} \|\mathsf{F}_n\nu - \tilde{\mathsf{F}}_n\nu\|_J &= \|\rho - \tilde{\rho}\|_{\{1\}\times J} \\ &\le 2(1-\varepsilon^{2\Delta}) \sum_{v\in J} \left\{ \sum_{v'\in(V\setminus K)\cup\partial K} D_{(1,v)(0,v')} + \sum_{v'\in V\setminus K} D_{(1,v)(1,v')} \right\} \\ &\le 4e^{-\beta}(1-\varepsilon^{2\Delta}) e^{-\beta d(J,\partial K)} \operatorname{card} J \end{aligned}$$

for every $J \subseteq K$. □

A.3 Decay of correlations of the block filter

To idea behind the block filter $\tilde{\pi}_n^\mu$ is that the error should decay exponentially in the block size by virtue of the decay of correlations property. While we have developed above the two ingredients (filter stability and one-step error bound) required to obtain a time-uniform error bound between π_n^μ and $\tilde{\pi}_n^\mu$, we have done this by imposing the decay of correlations property as an assumption. Thus perhaps the crucial point remains to be proved: we must show that decay of correlations does indeed hold, that is, that $\mathrm{Corr}(\tilde{\pi}_n^\mu, \beta)$ can be controlled uniformly in time. This is the goal of the present section.

Unfortunately, $\mathrm{Corr}(\tilde{\pi}_n^\mu, \beta)$ is not straightforward to control directly. We therefore introduce an alternative measure of correlation decay that will be easier to control. For any probability measure μ on $\mathbb{X}$ and $x, z \in \mathbb{X}$, $v \in V$, $K \in \mathcal{K}$, let

$$\begin{aligned} \mu_{x,z}^{v,K}(A) &:= \mathbf{P}^\mu(X_0^v \in A | X_0^{V\setminus\{v\}} = x^{V\setminus\{v\}},\ X_1^K = z^K) \\ &= \frac{\int \mathbf{1}_A(x^v) \prod_{w\in N(v)\cap K} p^w(x, z^w)\, \mu_x^v(dx^v)}{\int \prod_{w\in N(v)\cap K} p^w(x, z^w)\, \mu_x^v(dx^v)}. \end{aligned}$$

We now define

$$\tilde{C}_{vv'}^\mu := \frac{1}{2} \max_{K\in\mathcal{K}} \sup_{z\in\mathbb{X}} \sup_{x,\tilde{x}\in\mathbb{X}: x^{V\setminus\{v'\}} = \tilde{x}^{V\setminus\{v'\}}} \|\mu_{x,z}^{v,K} - \mu_{\tilde{x},z}^{v,K}\|$$

for $v, v' \in V$. The quantity

$$\widetilde{\mathrm{Corr}}(\mu, \beta) := \max_{v \in V} \sum_{v' \in V} e^{\beta d(v,v')} \tilde{C}^{\mu}_{vv'}$$

is a measure of correlation decay that is well adapted to the block filter. In order for this quantity to be useful, we must first show that it controls $\mathrm{Corr}(\mu, \beta)$.

Lemma A.8. *For any probability measure μ and $\beta > 0$, we have*

$$\mathrm{Corr}(\mu, \beta) \le (1 - \varepsilon^{2\Delta}) e^{2\beta r} \Delta^2 + 2\varepsilon^{-2\Delta} \widetilde{\mathrm{Corr}}(\mu, \beta).$$

Proof. By definition

$$\mu^v_{x,z}(A) = \frac{\int \mathbf{1}_A(x^v) \prod_{w \in N(v) \setminus K} p^w(x, z^w)\, \mu^{v,K}_{x,z}(dx^v)}{\int \prod_{w \in N(v) \setminus K} p^w(x, z^w)\, \mu^{v,K}_{x,z}(dx^v)}.$$

Let $x, \tilde{x} \in \mathbb{X}$ be such that $x^{V \setminus \{v'\}} = \tilde{x}^{V \setminus \{v'\}}$. If $v' \notin \bigcup_{w \in N(v)} N(w)$, then

$$\|\mu^v_{x,z} - \mu^v_{\tilde{x},z}\| \le 2\varepsilon^{-2\Delta} \|\mu^{v,K}_{x,z} - \mu^{v,K}_{\tilde{x},z}\|$$

by Lemma 2.9. On the other hand, note that

$$\mu^v_{x,z}(A) \ge \varepsilon^{2\Delta} \mu^{v,K}_{x,z}(A), \qquad \mu^v_{\tilde{x},z}(A) \ge \varepsilon^{2\Delta} \mu^{v,K}_{\tilde{x},z}(A).$$

We can therefore estimate using Lemma A.1 for $v' \in \bigcup_{w \in N(v)} N(w)$

$$\|\mu^v_{x,z} - \mu^v_{\tilde{x},z}\| \le 2(1 - \varepsilon^{2\Delta}) + \varepsilon^{2\Delta} \|\mu^{v,K}_{x,z} - \mu^{v,K}_{\tilde{x},z}\|.$$

Thus we obtain

$$\begin{aligned}\mathrm{Corr}(\mu, \beta) &\le (1 - \varepsilon^{2\Delta}) \max_{v \in V} \sum_{v' \in \bigcup_{w \in N(v)} N(w)} e^{\beta d(v,v')} + 2\varepsilon^{-2\Delta} \widetilde{\mathrm{Corr}}(\mu, \beta) \\ &\le (1 - \varepsilon^{2\Delta}) e^{2\beta r} \Delta^2 + 2\varepsilon^{-2\Delta} \widetilde{\mathrm{Corr}}(\mu, \beta).\end{aligned}$$

As μ and β were arbitrary, the proof is complete. □

We now aim to establish a time-uniform bound on $\widetilde{\mathrm{Corr}}(\tilde{\pi}^{\mu}_n, \beta)$. To this end, we first prove a one-step bound which will subsequently be iterated.

Proposition A.9. *Suppose there exists $\varepsilon > 0$ such that*

$$\varepsilon \le p^v(x, z^v) \le \varepsilon^{-1} \qquad \text{for all } v \in V,\ x, z \in \mathbb{X}.$$

Let ν be a probability measure on $\mathbb{X}$, and suppose that

$$\widetilde{\mathrm{Corr}}(\nu, \beta) + (1 - \varepsilon^2) e^{\beta(r+1)} \Delta \le \frac{1}{2}$$

for a sufficiently small constant $\beta > 0$. Then we have

$$\widetilde{\mathrm{Corr}}(\tilde{\mathsf{F}}_s \nu, \beta) \le 2(1 - \varepsilon^{2\Delta}) e^{2\beta r} \Delta^2$$

for any $s \in \mathbb{N}$.

Proof. Let $K, K' \in \mathcal{K}$, $v \in K$, $v' \in V$ $(v' \neq v)$, and let $z, x, \tilde{x} \in \mathbb{X}$ such that $x^{V\setminus\{v'\}} = \tilde{x}^{V\setminus\{v'\}}$. These choices will be fixed until further notice.

Define $I = (\{0\} \times V) \cup (1, v)$ and $\mathbb{S} = \mathbb{X} \times \mathbb{X}^v$, and let

$$\rho(A) = \frac{\int_A g^v(x^v, Y_s^v) \prod_{w\in K} p^w(x_0, x^w) \prod_{u\in N(v)\cap K'} p^u(x, z^u)\, \nu(dx_0)\, \psi^v(dx^v)}{\int g^v(x^v, Y_s^v) \prod_{w\in K} p^w(x_0, x^w) \prod_{u\in N(v)\cap K'} p^u(x, z^u)\, \nu(dx_0)\, \psi^v(dx^v)},$$

$$\tilde{\rho}(A) = \frac{\int_A g^v(\tilde{x}^v, Y_s^v) \prod_{w\in K} p^w(x_0, \tilde{x}^w) \prod_{u\in N(v)\cap K'} p^u(\tilde{x}, z^u)\, \nu(dx_0)\, \psi^v(d\tilde{x}^v)}{\int g^v(\tilde{x}^v, Y_s^v) \prod_{w\in K} p^w(x_0, \tilde{x}^w) \prod_{u\in N(v)\cap K'} p^u(\tilde{x}, z^u)\, \nu(dx_0)\, \psi^v(d\tilde{x}^v)},$$

Then we have by construction

$$\|(\tilde{\mathsf{F}}_s\nu)_{x,z}^{v,K'} - (\tilde{\mathsf{F}}_s\nu)_{\tilde{x},z}^{v,K'}\| = \|\rho - \tilde{\rho}\|_{(1,v)}.$$

We will apply Theorem 2.11 to bound $\|\rho - \tilde{\rho}\|_{(1,v)}$. To this end, we must bound C_{ij} and b_i with $i = (k, t)$ and $j = (k', t')$. We distinguish two cases.

Case $k = 0$. In this case we have

$$\rho^i_{(x_0,x^v)}(A) = \frac{\int \mathbf{1}_A(x_0^t) \prod_{w\in N(t)\cap K} p^w(x_0, x^w)\, \nu^t_{x_0}(dx_0^t)}{\int \prod_{w\in N(t)\cap K} p^w(x_0, x^w)\, \nu^t_{x_0}(dx_0^t)},$$

$$\tilde{\rho}^i_{(x_0,\tilde{x}^v)}(A) = \frac{\int \mathbf{1}_A(x_0^t) \prod_{w\in N(t)\cap K} p^w(x_0, \tilde{x}^w)\, \nu^t_{x_0}(dx_0^t)}{\int \prod_{w\in N(t)\cap K} p^w(x_0, \tilde{x}^w)\, \nu^t_{x_0}(dx_0^t)}.$$

Note that $\rho^i_{(x_0,x^v)} = \nu^{t,K}_{x_0,x}$. We therefore have $C_{ij} \leq \tilde{C}^{\nu}_{tt'}$ when $k' = 0$. Moreover,

$$\rho^i_{(x_0,x^v)}(A) \geq \varepsilon^2 \frac{\int \mathbf{1}_A(x_0^t) \prod_{w\in N(t)\cap(K\setminus\{v\})} p^w(x_0, x^w)\, \nu^t_{x_0}(dx_0^t)}{\int \prod_{w\in N(t)\cap(K\setminus\{v\})} p^w(x_0, x^w)\, \nu^t_{x_0}(dx_0^t)}$$

implies $C_{ij} \leq 1 - \varepsilon^2$ if $k' = 1$ and $v \in N(t)$ by Lemma A.1, and $C_{ij} = 0$ otherwise. On the other hand, note that as $x^{V\setminus\{v'\}} = \tilde{x}^{V\setminus\{v'\}}$ we have $\rho^i_{(x_0,x^v)} = \tilde{\rho}^i_{(x_0,x^v)}$ if $v' \notin N(t) \cap K$, while both $\rho^i_{(x_0,x^v)}(A)$ and $\tilde{\rho}^i_{(x_0,x^v)}(A)$ dominate

$$\varepsilon^2 \frac{\int \mathbf{1}_A(x_0^t) \prod_{w\in N(t)\cap(K\setminus\{v'\})} p^w(x_0, x^w)\, \nu^t_{x_0}(dx_0^t)}{\int \prod_{w\in N(t)\cap(K\setminus\{v'\})} p^w(x_0, x^w)\, \nu^t_{x_0}(dx_0^t)}.$$

Therefore, by Lemma A.1

$$b_{(0,t)} \leq \begin{cases} 0 & \text{for } v' \notin N(t) \cap K, \\ 2(1 - \varepsilon^2) & \text{otherwise.} \end{cases}$$

Case $k = 1$. In this case we have

$$\rho^i_{(x_0,x^v)}(A) = \frac{\int \mathbf{1}_A(x^v)\, g^v(x^v, Y_s^v)\, p^v(x_0, x^v) \prod_{u\in N(v)\cap K'} p^u(x, z^u)\, \psi^v(dx^v)}{\int g^v(x^v, Y_s^v)\, p^v(x_0, x^v) \prod_{u\in N(v)\cap K'} p^u(x, z^u)\, \psi^v(dx^v)},$$

$$\tilde{\rho}^i_{(x_0,\tilde{x}^v)}(A) = \frac{\int \mathbf{1}_A(\tilde{x}^v)\, g^v(\tilde{x}^v, Y_s^v)\, p^v(x_0, \tilde{x}^v) \prod_{u\in N(v)\cap K'} p^u(\tilde{x}, z^u)\, \psi^v(d\tilde{x}^v)}{\int g^v(\tilde{x}^v, Y_s^v)\, p^v(x_0, \tilde{x}^v) \prod_{u\in N(v)\cap K'} p^u(\tilde{x}, z^u)\, \psi^v(d\tilde{x}^v)}.$$

Estimating as above, we obtain $C_{ij} \le 1-\varepsilon^2$ whenever $k'=0$ and $t' \in N(v)$, and $C_{ij}=0$ otherwise. Similarly, arguing again as above, we obtain

$$b_{(1,v)} \le \begin{cases} 0 & \text{for } v' \notin \bigcup_{w \in N(v) \cap K'} N(w), \\ 2(1-\varepsilon^{2\Delta}) & \text{otherwise.} \end{cases}$$

Define the matrix $\{C_{ij}(v)\}_{i,j \in I}$ with the following entries:

$$\begin{aligned} C_{(0,t)(0,t')}(v) &= \tilde{C}^{\nu}_{tt'}, \\ C_{(0,t)(1,v)}(v) &= C_{(1,v)(0,t)}(v) = (1-\varepsilon^2)\mathbf{1}_{t \in N(v)}, \\ C_{(1,v)(1,v)}(v) &= 0. \end{aligned}$$

Combining the above two cases yields $C_{ij} \le C_{ij}(v)$, and we readily compute

$$\sum_{(k',t') \in I} e^{\beta\{|k-k'|+d(t,t')\}} C_{(k,t)(k',t')}(v) \le \widetilde{\mathrm{Corr}}(\nu,\beta) + (1-\varepsilon^2)e^{\beta(r+1)}\Delta \le \frac{1}{2}$$

where we have used the assumption of the Proposition. By Theorem 2.11

$$\begin{aligned} \|(\tilde{\mathsf{F}}_s\nu)^{v,K'}_{x,z} - (\tilde{\mathsf{F}}_s\nu)^{v,K'}_{\tilde{x},z}\| &= \|\rho - \tilde{\rho}\|_{(1,v)} \\ &\le 2(1-\varepsilon^2)\,\mathbf{1}_{v' \in K} \sum_{t' \in N(v')} D_{(1,v)(0,t')}(v) \\ &\quad + 2(1-\varepsilon^{2\Delta})\,\mathbf{1}_{v' \in \bigcup_{w \in N(v) \cap K'} N(w)}\, D_{(1,v)(1,v)}(v) \end{aligned}$$

where $D(v) := \sum_{n \ge 0} C(v)^n$. But note that the right-hand side does not depend on K' or $z, x, \tilde{x}$ (provided $x^{V \setminus \{v'\}} = \tilde{x}^{V \setminus \{v'\}}$). We therefore obtain

$$\begin{aligned} \tilde{C}^{\tilde{\mathsf{F}}_s\nu}_{vv'} &\le (1-\varepsilon^2)\,\mathbf{1}_{v' \in K} \sum_{t' \in N(v')} D_{(1,v)(0,t')}(v) \\ &\quad + (1-\varepsilon^{2\Delta})\,\mathbf{1}_{v' \in \bigcup_{w \in N(v) \cap K'} N(w)}\, D_{(1,v)(1,v)}(v) \end{aligned}$$

for every $K \in \mathcal{K}$, $v \in K$, and $v' \in V$.

To proceed, we note that

$$\begin{aligned} \sum_{v' \in V} e^{\beta d(v,v')} \tilde{C}^{\tilde{\mathsf{F}}_s\nu}_{vv'} &\le (1-\varepsilon^2) \sum_{v' \in K} e^{\beta d(v,v')} \sum_{t' \in N(v')} D_{(1,v)(0,t')}(v) \\ &\quad + (1-\varepsilon^{2\Delta})\, D_{(1,v)(1,v)}(v) \sum_{v' \in \bigcup_{w \in N(v) \cap K'} N(w)} e^{\beta d(v,v')} \\ &\le (1-\varepsilon^{2\Delta}) e^{2\beta r} \Delta^2 \sum_{(k',v') \in I} e^{\beta\{|1-k'|+d(v,v')\}} D_{(1,v)(k',v')}(v). \end{aligned}$$

Applying Lemma 2.13 to $C(v)$ yields the result. □

We now iterate the above result.

Corollary A.10. *Suppose there exists $\varepsilon > 0$ such that*

$$\varepsilon \le p^v(x, z^v) \le \varepsilon^{-1} \qquad \text{for all } v \in V,\ x, z \in \mathbb{X},$$

and such that

$$\varepsilon > \varepsilon_0 = \left(1 - \frac{1}{16\Delta^2}\right)^{1/2\Delta}.$$

Let μ be a probability measure on $\mathbb{X}$ such that

$$\widetilde{\mathrm{Corr}}(\mu, \beta) \le \frac{1}{8},$$

where $\beta = -(2r)^{-1} \log 16\Delta^2(1 - \varepsilon^{2\Delta}) > 0$. Then

$$\widetilde{\mathrm{Corr}}(\tilde{\pi}_n^\mu, \beta) \le \frac{1}{8} \qquad \text{for all } n \ge 0.$$

In particular, the latter holds whenever $\mu = \delta_x$ for any $x \in \mathbb{X}$.

Proof. The assumption $\varepsilon > \varepsilon_0$ implies $\beta > 0$ and

$$(1 - \varepsilon^2) e^{\beta(r+1)} \Delta \le \frac{1}{16}.$$

Therefore, if $\widetilde{\mathrm{Corr}}(\nu, \beta) \le 1/8$, then Proposition A.9 yields

$$\widetilde{\mathrm{Corr}}(\tilde{\mathsf{F}}_s \nu, \beta) \le 2(1 - \varepsilon^{2\Delta}) e^{2\beta r} \Delta^2 \le \frac{1}{8}.$$

Thus if $\widetilde{\mathrm{Corr}}(\mu, \beta) \le 1/8$, then $\widetilde{\mathrm{Corr}}(\tilde{\pi}_n^\mu, \beta) \le 1/8$ for all $n \ge 0$. Moreover, as $\widetilde{\mathrm{Corr}}(\delta_x, \beta) = 0$, the result hold automatically for $\mu = \delta_x$. □

We finally obtain the requisite bound on $\mathrm{Corr}(\tilde{\pi}_n^\mu, \beta)$ using Lemma A.8.

Corollary A.11 (Decay of correlations)**.** *Suppose there exists $\varepsilon > 0$ with*

$$\varepsilon \le p^v(x, z^v) \le \varepsilon^{-1} \qquad \text{for all } v \in V,\ x, z \in \mathbb{X},$$

such that

$$\varepsilon > \varepsilon_0 = \left(1 - \frac{1}{16\Delta^2}\right)^{1/2\Delta}.$$

Let $\beta = -(2r)^{-1} \log 16\Delta^2(1 - \varepsilon^{2\Delta}) > 0$. Then

$$\mathrm{Corr}(\tilde{\pi}_n^x, \beta) \le \frac{1}{3}$$

for every $n \ge 0$ and $x \in \mathbb{X}$.

Proof. By Corollary A.10 and Lemma A.8, we can estimate

$$\mathrm{Corr}(\tilde{\pi}_n^x, \beta) \le \frac{1}{16} + \frac{1}{4}\varepsilon^{-2\Delta} \le \frac{1}{3}$$

where we used that $\varepsilon^{2\Delta} \ge 1 - 1/16$. □

A.4 Bounding the bias

In the previous sections, we have proved a local filter stability bound (Proposition A.2), a local one-step error bound (Propositions A.6 and A.7), and decay of correlations of the block filter (Corollary A.11). We can now combine these results to obtain a time-uniform error bound between the filter and the block filter; this controls the bias of the block particle filtering algorithm.

Theorem A.12 (Bias term)**.** *Suppose there exists $\varepsilon > 0$ such that*

$$\varepsilon \le p^v(x, z^v) \le \varepsilon^{-1} \qquad \text{for all } v \in V,\ x, z \in \mathbb{X},$$

and such that

$$\varepsilon > \varepsilon_0 = \left(1 - \frac{1}{18\Delta^2}\right)^{1/2\Delta}.$$

Let $\beta = -(2r)^{-1} \log 18\Delta^2(1 - \varepsilon^{2\Delta}) > 0$. Then

$$\|\pi_n^x - \tilde{\pi}_n^x\|_J \le \frac{8e^{-\beta}}{1 - e^{-\beta}} (1 - \varepsilon^{2\Delta}) \operatorname{card} J\, e^{-\beta d(J, \partial K)}$$

for every $n \ge 0$, $x \in \mathbb{X}$, $K \in \mathcal{K}$ and $J \subseteq K$.

Proof. We begin with the elementary error decomposition

$$\|\pi_n^x - \tilde{\pi}_n^x\|_J \le \sum_{s=1}^{n} \|\mathsf{F}_n \cdots \mathsf{F}_{s+1} \mathsf{F}_s \tilde{\pi}_{s-1}^x - \mathsf{F}_n \cdots \mathsf{F}_{s+1} \tilde{\mathsf{F}}_s \tilde{\pi}_{s-1}^x\|_J.$$

We will bound each term in the sum.

Case $s = n$. To bound this term, note that

$$\operatorname{Corr}(\tilde{\pi}_{n-1}^x, \beta) + (1 - \varepsilon^2) e^{\beta(r+1)} \Delta \le \frac{1}{3} + \frac{1}{18} \le \frac{1}{2}$$

by Corollary A.11. Therefore, applying Proposition A.7 with $\nu = \tilde{\pi}_{n-1}^x$, we obtain

$$\|\mathsf{F}_n \tilde{\pi}_{n-1}^x - \tilde{\mathsf{F}}_n \tilde{\pi}_{n-1}^x\|_J \le 4e^{-\beta}(1 - \varepsilon^{2\Delta})\, e^{-\beta d(J, \partial K)} \operatorname{card} J.$$

Case $s < n$. To bound this term, note that by Corollary A.11

$$\operatorname{Corr}(\tilde{\pi}_s^x, \beta) + 3(1 - \varepsilon^{2\Delta}) e^{2\beta r} \Delta^2 \le \frac{1}{3} + \frac{1}{6} = \frac{1}{2}.$$

Applying Proposition A.2 with $\mu = \tilde{\pi}_s^x$ and $\nu = \mathsf{F}_s \tilde{\pi}_{s-1}^x$ yields

$$\begin{aligned} &\|\mathsf{F}_n \cdots \mathsf{F}_{s+1} \mathsf{F}_s \tilde{\pi}_{s-1}^x - \mathsf{F}_n \cdots \mathsf{F}_{s+1} \tilde{\mathsf{F}}_s \tilde{\pi}_{s-1}^x\|_J \\ &\quad \le 2e^{-\beta(n-s)} \sum_{v \in J} \max_{v' \in V} e^{-\beta d(v, v')} \sup_{x, z \in \mathbb{X}} \|(\mathsf{F}_s \tilde{\pi}_{s-1}^x)_{x,z}^{v'} - (\tilde{\mathsf{F}}_s \tilde{\pi}_{s-1}^x)_{x,z}^{v'}\|. \end{aligned}$$

On the other hand, as by Corollary A.11

$$\mathrm{Corr}(\tilde{\pi}^x_{s-1},\beta) + (1-\varepsilon^2)e^{\beta(r+1)}\Delta \le \frac{1}{3} + \frac{1}{18} \le \frac{1}{2},$$

we have by Proposition A.6 with $\nu = \tilde{\pi}^x_{s-1}$

$$\sup_{x,z\in\mathbb{X}} \|(\mathsf{F}_s\tilde{\pi}^x_{s-1})^{v'}_{x,z} - (\tilde{\mathsf{F}}_s\tilde{\pi}^x_{s-1})^{v'}_{x,z}\| \le 4e^{-\beta}(1-\varepsilon^{2\Delta})\, e^{-\beta d(v',\partial K)}.$$

We therefore obtain the estimate

$$\begin{aligned}\|\mathsf{F}_n\cdots\mathsf{F}_{s+1}\mathsf{F}_s\tilde{\pi}^x_{s-1} - \mathsf{F}_n\cdots\mathsf{F}_{s+1}\tilde{\mathsf{F}}_s\tilde{\pi}^x_{s-1}\|_J \\ \le 8e^{-\beta}(1-\varepsilon^{2\Delta})\, e^{-\beta(n-s)}e^{-\beta d(J,\partial K)}\,\mathrm{card}\, J,\end{aligned}$$

where we have used $d(v,v') + d(v',\partial K) \ge d(v,\partial K)$.

Substituting the above two cases into the error decomposition and summing the geometric series yields the statement of the Theorem. □

A.5 Local stability of the block filter

As was explained in Section 4.5.4, the chief difficulty in obtaining a time-uniform bound on the variance term is to establish stability of the block filter. This will be done in the present section.

We first establish a stability bound for nonrandom initial conditions.

Proposition A.13. *Suppose there exists $\varepsilon > 0$ such that*

$$\varepsilon \le p^v(x,z^v) \le \varepsilon^{-1} \qquad \text{for all } v\in V,\ x,z\in\mathbb{X},$$

and such that

$$\varepsilon > \varepsilon_0 = \left(1 - \frac{1}{6\Delta^2}\right)^{1/2\Delta}.$$

Let $\beta = -\log 6\Delta^2(1-\varepsilon^{2\Delta}) > 0$. Then

$$\|\tilde{\mathsf{F}}_n\cdots\tilde{\mathsf{F}}_{s+1}\delta_z - \tilde{\mathsf{F}}_n\cdots\tilde{\mathsf{F}}_{s+1}\delta_{z'}\|_J \le 4\,\mathrm{card}\, J\, e^{-\beta(n-s)}$$

for every $s<n$, $z,z'\in\mathbb{X}$, $K\in\mathcal{K}$, and $J\subseteq K$.

Proof. Fix throughout the proof $n>0$, $K\in\mathcal{K}$, and $J\subseteq K$. We will also assume throughout the proof for notational simplicity that $s=0$ (the ultimate conclusion will extend to any $s<n$ as in the proof of Proposition A.2).

We begin by constructing the computation tree as explained in section 4.5.4. For future reference, let us work first in the more general setting where the initial distributions $\mu = \bigotimes_{K'\in\mathcal{K}}\mu^{K'}$ and $\nu = \bigotimes_{K'\in\mathcal{K}}\nu^{K'}$ are independent across the blocks (rather than the special case of point masses δ_x and $\delta_{x'}$). Define for $K'\in\mathcal{K}$

$$N(K') = \{K''\in\mathcal{K} : d(K',K'') \le r\},$$

that is, $N(K')$ is the collection of blocks that interact with block K' in one step of the dynamics (recall that $\operatorname{card} N(K') \le \Delta_{\mathbb{K}}$). Then we can evidently write

$$\mathsf{B}^{K'}\tilde{\mathsf{F}}_s\mu = \mathsf{C}_s^{K'}\mathsf{P}^{K'} \bigotimes_{K''\in N(K')} \mu^{K''},$$

where we have defined for any probability η on $\mathbb{X}^{K'}$

$$(\mathsf{C}_s^{K'}\eta)(A) := \frac{\int \mathbf{1}_A(x^{K'}) \prod_{v\in K'} g^v(x^v, Y_s^v)\, \eta(dx^{K'})}{\int \prod_{v\in K'} g^v(x^v, Y_s^v)\, \eta(dx^{K'})},$$

and for any probability η on $\mathbb{X}^{\bigcup_{K''\in N(K')} K''}$

$$(\mathsf{P}^{K'}\eta)(A) := \int \mathbf{1}_A(x^{K'}) \prod_{v\in K'} p^v(z, x^v)\, \psi^v(dx^v)\, \eta(dz).$$

We therefore have

$$\mathsf{B}^K\tilde{\mathsf{F}}_n\cdots\tilde{\mathsf{F}}_1\mu =$$
$$\mathsf{C}_n^K\mathsf{P}^K \bigotimes_{K_{n-1}\in N(K)} \Bigg[\mathsf{C}_{n-1}^{K_{n-1}}\mathsf{P}^{K_{n-1}} \bigotimes_{K_{n-2}\in N(K_{n-1})} \Bigg[\mathsf{C}_{n-2}^{K_{n-2}}\mathsf{P}^{K_{n-2}}\cdots \bigotimes_{K_1\in N(K_2)} \Bigg[\mathsf{C}_1^{K_1}\mathsf{P}^{K_1} \bigotimes_{K_0\in N(K_1)} \mu^{K_0}\Bigg]\cdots\Bigg]\Bigg].$$

The structure of the computation tree is now readily visible in this expression. To formalize the construction, we introduce the tree index set

$$T := \{[K_u\cdots K_{n-1}] : 0 \le u < n,\ K_s \in N(K_{s+1}) \text{ for } u \le s < n\} \cup \{[\varnothing]\}$$

where we write $K_n := K$ for simplicity (recall that K and n are fixed throughout). The root of the tree $[\varnothing]$ represents the block K at time n, while $[K_u\cdots K_{n-1}]$ represents the duplicate of block K_u at time u that affects block K at time n along the branch $K_u \to K_{u+1} \to \cdots \to K_{n-1} \to K$ (cf. Figure 4.4 for a simple illustration). The vertex set corresponding to the computation tree is defined as

$$I = \{[K_u\cdots K_{n-1}]v : [K_u\cdots K_{n-1}] \in T,\ v \in K_u\} \cup \{[\varnothing]v : v \in K\},$$

and the corresponding state space is given by

$$\mathbb{S} = \prod_{i\in I} \mathbb{X}^i, \qquad \mathbb{X}^{[t]v} = \mathbb{X}^v \quad \text{for} \quad [t]v \in I.$$

It will be convenient in the sequel to introduce some additional notation. First, we will specify the children $c(i)$ of an index $i \in I$ as follows:

$$c([K_u\cdots K_{n-1}]v) := \{[K_{u-1}\cdots K_{n-1}]v' : K_{u-1} \in N(K_u),\ v' \in N(v)\},$$

and similarly for $c([\varnothing]v)$. Denote the depth $d(i)$ and location $v(i)$ of $i \in I$ as

$$d([K_u \cdots K_{n-1}]v) := u, \qquad d([\varnothing]v) := n, \qquad v([t]v) := v.$$

We define the index set of non-leaf vertices in I as

$$I_+ := \{i \in I : 0 < d(i) \le n\},$$

and the set of leaves of the tree T as

$$T_0 := \{[K_0 \cdots K_{n-1}] : K_s \in N(K_{s+1}) \text{ for } 0 \le s < n\}.$$

Finally, it will be natural to identify $[t] \in T$ with the corresponding subset of I:

$$[K_u \cdots K_{n-1}] = \{[K_u \cdots K_{n-1}]v : v \in K_u\},$$

together with the analogous identification for $[\varnothing]$.

We now define the probability measures $\rho, \tilde\rho$ on $\mathbb{S}$ as follows:

$$\rho(A) = \frac{\int \mathbf{1}_A(x) \prod_{i\in I_+} p^{v(i)}(x^{c(i)}, x^i)\, g^{v(i)}(x^i, Y^{v(i)}_{d(i)})\, \psi^{v(i)}(dx^i) \prod_{[t]\in T_0} \mu^{[t]}(dx^{[t]})}{\int \prod_{i\in I_+} p^{v(i)}(x^{c(i)}, x^i)\, g^{v(i)}(x^i, Y^{v(i)}_{d(i)})\, \psi^{v(i)}(dx^i) \prod_{[t]\in T_0} \mu^{[t]}(dx^{[t]})},$$

$$\tilde\rho(A) = \frac{\int \mathbf{1}_A(x) \prod_{i\in I_+} p^{v(i)}(x^{c(i)}, x^i)\, g^{v(i)}(x^i, Y^{v(i)}_{d(i)})\, \psi^{v(i)}(dx^i) \prod_{[t]\in T_0} \nu^{[t]}(dx^{[t]})}{\int \prod_{i\in I_+} p^{v(i)}(x^{c(i)}, x^i)\, g^{v(i)}(x^i, Y^{v(i)}_{d(i)})\, \psi^{v(i)}(dx^i) \prod_{[t]\in T_0} \nu^{[t]}(dx^{[t]})},$$

where we write $\mu^{[K_0\cdots K_{n-1}]} := \mu^{K_0}$ and $\nu^{[K_0\cdots K_{n-1}]} := \nu^{K_0}$ for simplicity. Then, by construction, the measure $\mathsf{B}^K\tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_1\mu$ coincides with the marginal of ρ on the root of the computation tree, while $\mathsf{B}^K\tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_1\nu$ coincides with the marginal of $\tilde\rho$ on the root of the computation tree. In particular, we obtain

$$\|\tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_1\mu - \tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_1\nu\|_J = \|\rho - \tilde\rho\|_{[\varnothing]J}.$$

We will use Theorem 2.11 to obtain a bound on this expression.

Throughout the remainder of the proof, we specialize to the case that $\mu = \delta_z$ and $\nu = \delta_{z'}$. To apply Theorem 2.11, we must bound the quantities C_{ij} and b_i with $i = [K_u \cdots K_{n-1}]v$ and $j = [K'_{u'} \cdots K'_{n-1}]v'$. We distinguish three cases.

Case $u = 0$. As $\mu = \delta_z$ is nonrandom we evidently have $\rho^i_x = \delta_{z^v}$, so that $C_{ij} = 0$. On the other hand, as $\tilde\rho^i_x = \delta_{z'^v}$, we cannot do better than $b_i \le 2$.

Case $0 < u < n$. Now we have

$$\rho^i_x(A) = \tilde\rho^i_x(A) = \frac{\int \mathbf{1}_A(x^i)\, g^v(x^i, Y^v_u)\, p^v(x^{c(i)}, x^i) \prod_{\ell\in I_+ : i\in c(\ell)} p^{v(\ell)}(x^{c(\ell)}, x^\ell)\, \psi^v(dx^i)}{\int g^v(x^i, Y^v_u)\, p^v(x^{c(i)}, x^i) \prod_{\ell\in I_+ : i\in c(\ell)} p^{v(\ell)}(x^{c(\ell)}, x^\ell)\, \psi^v(dx^i)}.$$

Thus $b_i = 0$. Moreover, by inspection, ρ_x^i does not depend on x^j except in the following cases: $j \in c(i)$; $i \in c(j)$; $j \in c(\ell)$ for some $\ell \in I_+$ such that $i \in c(\ell)$. As $\operatorname{card} c(\ell) \le \Delta$ for every $\ell \in I_+$, we estimate using Lemma A.1

$$C_{ij} \le \begin{cases} 1-\varepsilon^2 & \text{if } j \in c(i), \\ 1-\varepsilon^2 & \text{if } i \in c(j), \\ 1-\varepsilon^{2\Delta} & \text{if } j \in \bigcup_{\ell \in I_+ : i \in c(\ell)} c(\ell), \\ 0 & \text{otherwise.} \end{cases}$$

This yields

$$\sum_{j \in I} e^{\beta|d(i)-d(j)|} C_{ij} \le 2(1-\varepsilon^2)e^\beta \Delta + (1-\varepsilon^{2\Delta})\Delta^2 \le 3(1-\varepsilon^{2\Delta})e^\beta \Delta^2,$$

where we have used that $\beta > 0$ and $\Delta \ge 1$ in the last inequality.

Case $u = n$. Now $i = [\varnothing]v$, so we have

$$\rho_x^i(A) = \tilde\rho_x^i(A) = \frac{\int \mathbf{1}_A(x^i)\, g^v(x^i, Y_n^v)\, p^v(x^{c(i)}, x^i)\, \psi^v(dx^i)}{\int g^v(x^i, Y_n^v)\, p^v(x^{c(i)}, x^i)\, \psi^v(dx^i)}.$$

Arguing precisely as above, we obtain $b_i = 0$ and

$$\sum_{j \in I} e^{\beta|d(i)-d(j)|} C_{ij} \le (1-\varepsilon^2)e^\beta \Delta.$$

Combining the above three cases, we obtain

$$\max_{i \in I} \sum_{j \in I} e^{\beta|d(i)-d(j)|} C_{ij} \le 3(1-\varepsilon^{2\Delta})e^\beta \Delta^2 = \frac{1}{2}$$

by the assumption of the Proposition. Thus by Theorem 2.11

$$\|\tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_1 \delta_z - \tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_1 \delta_{z'}\|_J = \|\rho - \tilde\rho\|_{[\varnothing]J} \le 4 \operatorname{card} J\, e^{-\beta n},$$

where we have used Lemma 2.13 with $m(i,j) = \beta|d(i)-d(j)|$. The proof is completed by extending to general $s < n$ as in the proof of Proposition A.2. □

The proof of Proposition A.13 was simplified by the fact that the resulting bound holds uniformly for all point mass initial conditions (this could be used to obtain a uniform bound for all initial measures along the same lines as the proof of Corollary A.5). To obtain a bound on the variance term, however, we require a more precise stability bound for the block filter that provides explicit control in terms of the initial conditions. We will shortly deduce such a bound from Proposition A.13. Before we can do so, however, we must prove a refinement of Lemma 2.9.

Lemma A.14. *Let $\mu = \mu^1 \otimes \cdots \otimes \mu^d$ and $\nu = \nu^1 \otimes \cdots \otimes \nu^d$ be product probability measures on $\mathbb{S} = \mathbb{S}^1 \times \cdots \times \mathbb{S}^d$, and let $\Lambda : \mathbb{S} \to \mathbb{R}$ be a bounded and strictly positive measurable function. Define the probability measures*

$$\mu_\Lambda(A) := \frac{\int \mathbf{1}_A(x)\Lambda(x)\mu(dx)}{\int \Lambda(x)\mu(dx)}, \qquad \nu_\Lambda(A) := \frac{\int \mathbf{1}_A(x)\Lambda(x)\nu(dx)}{\int \Lambda(x)\nu(dx)}.$$

Suppose that there exists a constant $\varepsilon > 0$ such that the following holds: for every $i = 1, \ldots, d$, there is a measurable function $\Lambda^i : \mathbb{S} \to \mathbb{R}$ such that

$$\varepsilon\Lambda^i(x) \le \Lambda(x) \le \varepsilon^{-1}\Lambda^i(x) \quad \text{for all } x \in \mathbb{S}$$

and such that $\Lambda^i(x) = \Lambda^i(\tilde{x})$ whenever $x^{\{1,\ldots,d\}\setminus\{i\}} = \tilde{x}^{\{1,\ldots,d\}\setminus\{i\}}$. Then

$$\|\mu_\Lambda - \nu_\Lambda\| \le \frac{2}{\varepsilon^2}\sum_{i=1}^d \|\mu^i - \nu^i\|.$$

Proof. Define for $i = 0, \ldots, d$ the measures

$$\rho_i := \nu^1 \otimes \cdots \otimes \nu^i \otimes \mu^{i+1} \otimes \cdots \otimes \mu^d, \qquad \rho_{i,\Lambda}(A) := \frac{\int \mathbf{1}_A(x)\Lambda(x)\rho_i(dx)}{\int \Lambda(x)\rho_i(x)}$$

(by convention, $\rho_0 = \mu$ and $\rho_d = \nu$). Then we can estimate

$$\|\mu_\Lambda - \nu_\Lambda\| \le \sum_{i=1}^d \|\rho_{i,\Lambda} - \rho_{i-1,\Lambda}\|.$$

Now note that we can estimate for $|f| \le 1$

$$|\rho_{i,\Lambda}(f) - \rho_{i-1,\Lambda}(f)| \le \frac{1}{\varepsilon\rho_i(\Lambda^i)}\Big[|\rho_i(f\Lambda) - \rho_{i-1}(f\Lambda)| + |\rho_i(\Lambda) - \rho_{i-1}(\Lambda)|\Big]$$

as in the proof of Lemma 2.9. Moreover, we can write

$$|\rho_i(f\Lambda) - \rho_{i-1}(f\Lambda)| = \frac{\rho_i(\Lambda^i)}{\varepsilon}\left|\int f^i(x)\nu^i(dx^i) - \int f^i(x)\mu^i(dx^i)\right|,$$
$$|\rho_i(\Lambda) - \rho_{i-1}(\Lambda)| = \frac{\rho_i(\Lambda^i)}{\varepsilon}\left|\int g^i(x)\nu^i(dx^i) - \int g^i(x)\mu^i(dx^i)\right|,$$

where f^i and g^i are functions on $\mathbb{S}^i$ defined by

$$f^i(x^i) := \frac{\varepsilon}{\rho_i(\Lambda^i)}\int f(x)\Lambda(x)\,\nu^1(dx^1)\cdots\nu^{i-1}(dx^{i-1})\,\mu^{i+1}(dx^{i+1})\cdots\mu^d(dx^d),$$
$$g^i(x^i) := \frac{\varepsilon}{\rho_i(\Lambda^i)}\int \Lambda(x)\,\nu^1(dx^1)\cdots\nu^{i-1}(dx^{i-1})\,\mu^{i+1}(dx^{i+1})\cdots\mu^d(dx^d).$$

Evidently $|f^i| \le 1$ and $|g^i| \le 1$, and the proof follows directly. □

We can now obtain a stability bound with control on the initial conditions.

Proposition A.15. *Suppose there exists $\varepsilon > 0$ with*

$$\varepsilon \le p^v(x, z^v) \le \varepsilon^{-1} \qquad \text{for all } v \in V,\ x, z \in \mathbb{X}$$

such that

$$\varepsilon > \varepsilon_0 = \left(1 - \frac{1}{6\Delta^2}\right)^{1/2\Delta}.$$

Let $\beta = -\log 6\Delta^2(1 - \varepsilon^{2\Delta}) > 0$. Then for any product probability measures

$$\mu = \bigotimes_{K \in \mathcal{K}} \mu^K, \qquad \nu = \bigotimes_{K \in \mathcal{K}} \nu^K,$$

we have

$$\|\tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_{s+1}\mu - \tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_{s+1}\nu\|_J \le \frac{4}{\varepsilon^{2|\mathcal{K}|_\infty}} \operatorname{card} J\, e^{-\beta(n-s)} \sum_{K \in \mathcal{K}} \alpha_K \|\mu^K - \nu^K\|$$

for every $s < n$, $K \in \mathcal{K}$, and $J \subseteq K$. Here $(\alpha_K)_{K \in \mathcal{K}}$ are nonnegative integers, depending on J and $n - s$ only, such that $\sum_{K \in \mathcal{K}} \alpha_K \le \Delta_{\mathcal{K}}^{n-s}$.

Proof. We fix $s = 0$, $n > 0$, $K \in \mathcal{K}$, $J \subseteq K$ as in the proof of Proposition A.13, and adopt the notation used there. Define the functions

$$h_A(x^{T_0}) := \int \mathbf{1}_A(x^{[\varnothing]J}) \prod_{i \in I_+} p^{v(i)}(x^{c(i)}, x^i)\, g^{v(i)}(x^i, Y^{v(i)}_{d(i)})\, \psi^{v(i)}(dx^i),$$

$$h(x^{T_0}) := \int \prod_{i \in I_+} p^{v(i)}(x^{c(i)}, x^i)\, g^{v(i)}(x^i, Y^{v(i)}_{d(i)})\, \psi^{v(i)}(dx^i)$$

on the leaves T_0 of the computation tree, for every measurable $A \subseteq \mathbb{X}^J$. Then

$$(\tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_1\mu)(A) = \frac{\int h_A(x^{T_0}) \prod_{[t] \in T_0} \mu^{[t]}(dx^{[t]})}{\int h(x^{T_0}) \prod_{[t] \in T_0} \mu^{[t]}(dx^{[t]})} = \int \frac{h_A(x^{T_0})}{h(x^{T_0})}\, \tilde{\mu}(dx^{T_0}),$$

where we define the measure

$$\tilde{\mu}(A) := \frac{\int \mathbf{1}_A(x^{T_0})\, h(x^{T_0}) \prod_{[t] \in T_0} \mu^{[t]}(dx^{[t]})}{\int h(x^{T_0}) \prod_{[t] \in T_0} \mu^{[t]}(dx^{[t]})}.$$

The measure $\tilde{\nu}$ is define analogously, and we have

$$\|\tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_1\mu - \tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_1\nu\|_J = 2 \sup_{A \subseteq \mathbb{X}^J} \left| \int \frac{h_A}{h}\, d\tilde{\mu} - \int \frac{h_A}{h}\, d\tilde{\nu} \right|,$$

where the supremum is taken only over measurable sets. But note that h_A/h is precisely the filter obtained when the initial condition is a point mass on the leaves

of the computation tree (albeit not with the special duplication pattern induced by the unravelling of the original model; however, this was not used in the proof of Proposition A.13). Therefore, the proof of Proposition A.13 yields

$$2 \sup_{z,\tilde{z}\in\mathbb{X}^{T_0}} \sup_{A\subseteq\mathbb{X}^J} \left| \frac{h_A(z)}{h(z)} - \frac{h_A(\tilde{z})}{h(\tilde{z})} \right| \le 4 \operatorname{card} J \, e^{-\beta n}.$$

In particular, using the identity $|\mu(f) - \nu(f)| \le \frac{1}{2} \operatorname{osc} f \, \|\mu - \nu\|$, we obtain

$$\|\tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_1 \mu - \tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_1 \nu\|_J \le 2 \operatorname{card} J \, e^{-\beta n} \, \|\tilde{\mu} - \tilde{\nu}\|.$$

We now aim to apply Lemma A.14 to estimate $\|\tilde{\mu} - \tilde{\nu}\|$.

To this end, consider a block $[t] \in T_0$. The integrand in the definition of $h(x^{T_0})$ depends only on $x^{[t]}$ through the terms $p^{v(i)}(x^{c(i)}, x^i)$ with $c(i) \cap [t] \neq \varnothing$. If we write $[t] = [K_0 \cdots K_{n-1}]$, then $c(i) \cap [t] \neq \varnothing$ requires at least $i \in [K_1 \cdots K_{n-1}]$ and therefore $\operatorname{card}\{i \in I_+ : c(i) \cap [t] \neq \varnothing\} \le \operatorname{card} K_1 \le |\mathcal{K}|_\infty$. Thus we have

$$\varepsilon^{|\mathcal{K}|_\infty} h^{[t]}(z) \le h(z) \le \varepsilon^{-|\mathcal{K}|_\infty} h^{[t]}(z)$$

for every $z \in \mathbb{X}^{T_0}$ and $[t] \in T_0$, where

$$h^{[t]}(x^{T_0}) := \int \prod_{i\in I_+ : c(i)\cap[t]=\varnothing} p^{v(i)}(x^{c(i)}, x^i) \prod_{i\in I_+} g^{v(i)}(x^i, Y^{v(i)}_{d(i)}) \, \psi^{v(i)}(dx^i)$$

does not depend on $x^{[t]}$. By Lemma A.14, we obtain

$$\|\tilde{\mu} - \tilde{\nu}\| \le \frac{2}{\varepsilon^{2|\mathcal{K}|_\infty}} \sum_{[t]\in T_0} \|\mu^{[t]} - \nu^{[t]}\| = \frac{2}{\varepsilon^{2|\mathcal{K}|_\infty}} \sum_{K'\in\mathcal{K}} \alpha_{K'} \|\mu^{K'} - \nu^{K'}\|,$$

where we define $\alpha_{K'} = \operatorname{card}\{[K_0 \cdots K_{n-1}] \in T_0 : K_0 = K'\}$. As the computation tree has a branching factor of at most $\Delta_{\mathcal{K}}$, we evidently have $\sum_{K\in\mathcal{K}} \alpha_K = \operatorname{card} T_0 \le \Delta_{\mathcal{K}}^n$. The result therefore follows directly for the case $s = 0$, and the general case $s < n$ is immediate as in the proof of Proposition A.2. □

We finally state the block filter stability bound in its most useful form.

Corollary A.16 (Block filter stability)**.** *Suppose there exists $\varepsilon > 0$ with*

$$\varepsilon \le p^v(x, z^v) \le \varepsilon^{-1} \qquad \text{for all } v \in V,\ x, z \in \mathbb{X}$$

such that

$$\varepsilon > \varepsilon_0 = \left(1 - \frac{1}{6\Delta_{\mathcal{K}}\Delta^2}\right)^{1/2\Delta}.$$

Let $\beta = -\log 6\Delta_{\mathcal{K}}\Delta^2(1 - \varepsilon^{2\Delta}) > 0$.

Then for any (possibly random) product probability measures

$$\mu = \bigotimes_{K\in\mathcal{K}} \mu^K, \qquad \nu = \bigotimes_{K\in\mathcal{K}} \nu^K,$$

we have

$$\sqrt{\mathbf{E}\,\|\tilde{\mathsf{F}}_n\cdots\tilde{\mathsf{F}}_{s+1}\mu-\tilde{\mathsf{F}}_n\cdots\tilde{\mathsf{F}}_{s+1}\nu\|_J^2}$$
$$\le\frac{4}{\varepsilon^{2|\mathcal{K}|_\infty}}\,\mathrm{card}\,J\,e^{-\beta(n-s)}\max_{K\in\mathcal{K}}\sqrt{\mathbf{E}\,\|\mu^K-\nu^K\|^2}$$

for every $s<n$, $K\in\mathcal{K}$, and $J\subseteq K$.

Proof. The result follows readily from Proposition A.15 (note that we have now absorbed the branching factor $\Delta_{\mathcal{K}}^{n-s}$ in the definition of β). □

A.6 Bounding the variance

To complete the proof of Theorem 4.2, it now remains to bound the variance term $|||\tilde{\pi}_n-\hat{\pi}_n|||_J$ uniformly in time. This is the goal of the present section. We will first obtain bounds on the one-step error, and then combine these with the block filter stability bound of Corollary A.16 to obtain time-uniform control of the error. The main remaining difficulty is to properly account for the fact that Corollary A.16 is phrased in terms of the total variation norm $\|\cdot\|_J$, which is too strong to control the sampling error (we do not know how to prove an analogous result to Corollary A.16 in the weaker $|||\cdot|||_J$-norm). To this end, we retain one time step of the block filter dynamics in the one-step error (we control $\|\tilde{\mathsf{F}}_{s+1}\tilde{\mathsf{F}}_s\hat{\pi}_{s-1}^x-\tilde{\mathsf{F}}_{s+1}\hat{\mathsf{F}}_s\hat{\pi}_{s-1}^x\|_K$ rather than $|||\tilde{\mathsf{F}}_s\hat{\pi}_{s-1}^x-\hat{\mathsf{F}}_s\hat{\pi}_{s-1}^x|||_K$), which allows us to exploit the fact that the dynamics P has a density.

Let us begin with the most trivial result: a one-step bound in the $|||\cdot|||_J$-norm. This estimate will be used to bound the error in the last time step $s=n$.

Lemma A.17 (Sampling error, $s=n$)**.** *Suppose there exists $\kappa>0$ such that*

$$\kappa\le g^v(x^v,y^v)\le\kappa^{-1}\qquad\text{for all } v\in V,\ x\in\mathbb{X},\ y\in\mathbb{Y}.$$

Then

$$\max_{K\in\mathcal{K}}|||\tilde{\mathsf{F}}_n\hat{\pi}_{n-1}^\mu-\hat{\mathsf{F}}_n\hat{\pi}_{n-1}^\mu|||_K\le\frac{2\kappa^{-2|\mathcal{K}|_\infty}}{\sqrt{N}}.$$

Proof. Note that

$$|||\tilde{\mathsf{F}}_n\hat{\pi}_{n-1}^\mu-\hat{\mathsf{F}}_n\hat{\pi}_{n-1}^\mu|||_K=|||\mathsf{C}_n^K\mathsf{B}^K\mathsf{P}\hat{\pi}_{n-1}^\mu-\mathsf{C}_n^K\mathsf{B}^K\mathsf{S}^N\mathsf{P}\hat{\pi}_{n-1}^\mu|||$$
$$\le2\kappa^{-2\,\mathrm{card}\,K}|||\mathsf{P}\hat{\pi}_{n-1}^\mu-\mathsf{S}^N\mathsf{P}\hat{\pi}_{n-1}^\mu|||\le\frac{2\kappa^{-2\,\mathrm{card}\,K}}{\sqrt{N}},$$

where the first inequality is Lemma 2.9 and the second inequality follows from the simple estimate $|||\mu-\mathsf{S}^N\mu|||\le1/\sqrt{N}$ that holds for any probability μ. □

For the error in steps $s<n$, the requisite one-step bound (Proposition A.20) is more involved. Before we prove it, we must first introduce an elementary lemma about products of empirical measures that will be needed below.

Lemma A.18. *For any probability measure μ, we have*

$$|||\mu^{\otimes d} - \hat{\mu}^{\otimes d}||| \le \frac{4d}{\sqrt{N}},$$

where $\hat{\mu} = \frac{1}{N}\sum_{k=1}^N \delta_{X_k}$ and $X_1, \ldots, X_N$ are i.i.d. $\sim \mu$.

Proof. We assume throughout that $N \ge d^2$ without loss of generality (otherwise the bound is trivial). Let $|f| \le 1$ be a measurable function. Then

$$\hat{\mu}^{\otimes d}(f) = \frac{1}{N^d} \sum_{k_1,\ldots,k_d=1}^N f(X_{k_1}, \ldots, X_{k_d}).$$

We begin by bounding

$$\mathrm{Var}[\hat{\mu}^{\otimes d}(f)] = \frac{1}{N^{2d}} \sum_{k_1,\ldots,k_d=1}^N \sum_{k'_1,\ldots,k'_d=1}^N \mathbf{E}(F_{k_1,\ldots,k_d} F_{k'_1,\ldots,k'_d})$$

where

$$F_{k_1,\ldots,k_d} := f(X_{k_1}, \ldots, X_{k_d}) - \mathbf{E}\, f(X_{k_1}, \ldots, X_{k_d}).$$

Note that $\mathbf{E}(F_{k_1,\ldots,k_d} F_{k'_1,\ldots,k'_d}) = 0$ when $\{k_1, \ldots, k_d\} \cap \{k'_1, \ldots, k'_d\} = \varnothing$. Thus

$$\mathrm{Var}[\hat{\mu}^{\otimes d}(f)] \le \frac{4}{N^{2d}} \sum_{k_1,\ldots,k_d=1}^N \sum_{k'_1,\ldots,k'_d=1}^N \mathbf{1}_{\{k_1,\ldots,k_d\} \cap \{k'_1,\ldots,k'_d\} \neq \varnothing},$$

where we use $|F_{k_1,\ldots,k_d}| \le 2$. But for each choice of $k_1, \ldots, k_d$, there are at least $(N-d)^d$ choices of $k'_1, \ldots, k'_d$ such that $\{k_1, \ldots, k_d\} \cap \{k'_1, \ldots, k'_d\} = \varnothing$, so

$$\mathrm{Var}[\hat{\mu}^{\otimes d}(f)] \le 4\left(1 - \frac{N^d (N-d)^d}{N^{2d}}\right) = 4\left(1 - \left(1 - \frac{d}{N}\right)^d\right) \le \frac{4d^2}{N}.$$

We can therefore estimate

$$\begin{aligned} |||\mu^{\otimes d} - \hat{\mu}^{\otimes d}||| &\le \|\mu^{\otimes d} - \mathbf{E}\,\hat{\mu}^{\otimes d}\| + |||\mathbf{E}\,\hat{\mu}^{\otimes d} - \hat{\mu}^{\otimes d}||| \\ &\le \|\mu^{\otimes d} - \mathbf{E}\,\hat{\mu}^{\otimes d}\| + \frac{2d}{\sqrt{N}}. \end{aligned}$$

It remains to estimate the first term. To this end, note that $\mathbf{E}\, f(X_{k_1}, \ldots, X_{k_n}) = \mu^{\otimes d}(f)$ whenever $k_1 \neq \cdots \neq k_n$. Therefore, we evidently have

$$\begin{aligned} |\mathbf{E}\,\hat{\mu}^{\otimes d}(f) - \mu^{\otimes d}(f)| &\le \frac{1}{N^d} \sum_{k_1,\ldots,k_d=1}^N |\mathbf{E}\, f(X_{k_1}, \ldots, X_{k_d}) - \mu^{\otimes d}(f)| \\ &\le 2\left(1 - \frac{1}{N^d}\frac{N!}{(N-d)!}\right) \le 2\left(1 - \left(1 - \frac{d}{N}\right)^d\right) \le \frac{2d^2}{N}. \end{aligned}$$

But as $N \ge d^2$, we have $d^2/N \le d/\sqrt{N}$. The result follows. □

This result will be used in the following form.

Corollary A.19. *For any subset of blocks $\mathcal{L} \subseteq \mathcal{K}$, we have*

$$\|\!|\bigotimes_{K\in\mathcal{L}} \mathsf{B}^K\mu - \bigotimes_{K\in\mathcal{L}} \mathsf{B}^K\mathsf{S}^N\mu|\!\| \le \frac{4\,\mathrm{card}\,\mathcal{L}}{\sqrt{N}}$$

for every probability measure μ on $\mathbb{X}$ and $s \ge 1$.

Proof. Write $\hat{\mu} := \mathsf{S}^N\mu$ and $d = \mathrm{card}\,\mathcal{L}$, and let us enumerate the blocks $\mathcal{L} = \{K_1, \dots, K_d\}$. Then for any bounded function $f : \mathbb{X}^{\bigcup\mathcal{L}} \to \mathbb{R}$, we can write

$$(\bigotimes_{K\in\mathcal{L}} \mathsf{B}^K\mu)(f) = \int f(x_1^{K_1}, \dots, x_d^{K_d})\,\mu(dx_1)\cdots\mu(dx_d),$$
$$(\bigotimes_{K\in\mathcal{L}} \mathsf{B}^K\mathsf{S}^N\mu)(f) = \int f(x_1^{K_1}, \dots, x_d^{K_d})\,\hat{\mu}(dx_1)\cdots\hat{\mu}(dx_d).$$

Thus evidently

$$\|\!|\bigotimes_{K\in\mathcal{L}} \mathsf{B}^K\mu - \bigotimes_{K\in\mathcal{L}} \mathsf{B}^K\mathsf{S}^N\mu|\!\| \le \|\!|\mu^{\otimes d} - \hat{\mu}^{\otimes d}|\!\|,$$

and the result follows from Lemma A.18. □

We now proceed to prove a one-step error bound for time steps $s < n$.

Proposition A.20 (Sampling error, $s < n$)**.** *Suppose there exist $\varepsilon, \kappa > 0$ with*

$$\varepsilon \le p^v(x, z^v) \le \varepsilon^{-1}, \quad \kappa \le g^v(x^v, y^v) \le \kappa^{-1} \quad \forall\, v \in V,\ x, z \in \mathbb{X},\ y \in \mathbb{Y}.$$

Then

$$\max_{K\in\mathcal{K}} \sqrt{\mathbf{E}\,\|\tilde{\mathsf{F}}_{s+1}\tilde{\mathsf{F}}_s\hat{\pi}^\mu_{s-1} - \tilde{\mathsf{F}}_{s+1}\hat{\mathsf{F}}_s\hat{\pi}^\mu_{s-1}\|_K^2} \le \frac{16\Delta_{\mathcal{K}}\,\varepsilon^{-2|\mathcal{K}|_\infty}\kappa^{-4|\mathcal{K}|_\infty\Delta_{\mathcal{K}}}}{\sqrt{N}}$$

for every $0 < s < n$.

Proof. We begin by bounding using Lemma 2.9

$$\begin{aligned}\|\tilde{\mathsf{F}}_{s+1}\tilde{\mathsf{F}}_s\hat{\pi}^\mu_{s-1} - \tilde{\mathsf{F}}_{s+1}\hat{\mathsf{F}}_s\hat{\pi}^\mu_{s-1}\|_K &= \|\mathsf{C}^K_{s+1}\mathsf{B}^K\mathsf{P}\tilde{\mathsf{F}}_s\hat{\pi}^\mu_{s-1} - \mathsf{C}^K_{s+1}\mathsf{B}^K\mathsf{P}\hat{\mathsf{F}}_s\hat{\pi}^\mu_{s-1}\| \\ &\le 2\kappa^{-2|\mathcal{K}|_\infty}\|\mathsf{B}^K\mathsf{P}\tilde{\mathsf{F}}_s\hat{\pi}^\mu_{s-1} - \mathsf{B}^K\mathsf{P}\hat{\mathsf{F}}_s\hat{\pi}^\mu_{s-1}\|.\end{aligned}$$

Now note that

$$\frac{(\mathsf{B}^K\mathsf{P}\tilde{\mathsf{F}}_s\hat{\pi}^\mu_{s-1})(dx^K)}{\psi^K(dx^K)} = \frac{\int \prod_{v\in K} p^v(z, x^v) \prod_{K'\in N(K)} \prod_{v'\in K'} g^{v'}(z^{v'}, Y_s^{v'})(\mathsf{B}^{K'}\mathsf{P}\hat{\pi}^\mu_{s-1})(dz^{K'})}{\int \prod_{K'\in N(K)} \prod_{v'\in K'} g^{v'}(z^{v'}, Y_s^{v'})(\mathsf{B}^{K'}\mathsf{P}\hat{\pi}^\mu_{s-1})(dz^{K'})},$$
$$\frac{(\mathsf{B}^K\mathsf{P}\hat{\mathsf{F}}_s\hat{\pi}^\mu_{s-1})(dx^K)}{\psi^K(dx^K)} = \frac{\int \prod_{v\in K} p^v(z, x^v) \prod_{K'\in N(K)} \prod_{v'\in K'} g^{v'}(z^{v'}, Y_s^{v'})(\mathsf{B}^{K'}\mathsf{S}^N\mathsf{P}\hat{\pi}^\mu_{s-1})(dz^{K'})}{\int \prod_{K'\in N(K)} \prod_{v'\in K'} g^{v'}(z^{v'}, Y_s^{v'})(\mathsf{B}^{K'}\mathsf{S}^N\mathsf{P}\hat{\pi}^\mu_{s-1})(dz^{K'})},$$

where $\psi^K(dx^K) := \prod_{v\in K}\psi^v(dx^v)$, and we can write

$$\|\mathsf{B}^K\mathsf{P}\tilde{\mathsf{F}}_s\hat{\pi}^\mu_{s-1} - \mathsf{B}^K\mathsf{P}\hat{\mathsf{F}}_s\hat{\pi}^\mu_{s-1}\| = \int\left|\frac{(\mathsf{B}^K\mathsf{P}\tilde{\mathsf{F}}_s\hat{\pi}^\mu_{s-1})(dx^K)}{\psi^K(dx^K)} - \frac{(\mathsf{B}^K\mathsf{P}\hat{\mathsf{F}}_s\hat{\pi}^\mu_{s-1})(dx^K)}{\psi^K(dx^K)}\right|\psi^K(dx^K).$$

We therefore have by Minkowski's integral inequality

$$\begin{aligned}&\sqrt{\mathbf{E}\,\|\mathsf{B}^K\mathsf{P}\tilde{\mathsf{F}}_s\hat{\pi}^\mu_{s-1} - \mathsf{B}^K\mathsf{P}\hat{\mathsf{F}}_s\hat{\pi}^\mu_{s-1}\|^2}\\ &\le \int\sqrt{\mathbf{E}\left|\frac{(\mathsf{B}^K\mathsf{P}\tilde{\mathsf{F}}_s\hat{\pi}^\mu_{s-1})(dx^K)}{\psi^K(dx^K)} - \frac{(\mathsf{B}^K\mathsf{P}\hat{\mathsf{F}}_s\hat{\pi}^\mu_{s-1})(dx^K)}{\psi^K(dx^K)}\right|^2}\psi^K(dx^K)\\ &\le \psi^K(\mathbb{X}^K)\sup_{x^K\in\mathbb{X}^K}\sqrt{\mathbf{E}\left|\frac{(\mathsf{B}^K\mathsf{P}\tilde{\mathsf{F}}_s\hat{\pi}^\mu_{s-1})(dx^K)}{\psi^K(dx^K)} - \frac{(\mathsf{B}^K\mathsf{P}\hat{\mathsf{F}}_s\hat{\pi}^\mu_{s-1})(dx^K)}{\psi^K(dx^K)}\right|^2}.\end{aligned}$$

As we have

$$\varepsilon\psi^v(\mathbb{X}^v) \le \int p^v(x,z^v)\psi^v(dz^v) = 1, \qquad \prod_{v\in K} p^v(z,x^v) \le \varepsilon^{-|\mathcal{K}|_\infty},$$

and

$$\kappa^{|\mathcal{K}|_\infty\Delta_\mathcal{K}} \le \prod_{K'\in N(K)}\prod_{v'\in K'} g^{v'}(z^{v'},Y^{v'}_s) \le \kappa^{-|\mathcal{K}|_\infty\Delta_\mathcal{K}},$$

we can apply Lemma 2.9 to estimate

$$\begin{aligned}&\sqrt{\mathbf{E}\,\|\mathsf{B}^K\mathsf{P}\tilde{\mathsf{F}}_s\hat{\pi}^\mu_{s-1} - \mathsf{B}^K\mathsf{P}\hat{\mathsf{F}}_s\hat{\pi}^\mu_{s-1}\|^2}\\ &\quad\le 2\varepsilon^{-2|\mathcal{K}|_\infty}\kappa^{-2|\mathcal{K}|_\infty\Delta_\mathcal{K}}\|\!|\!|\bigotimes_{K'\in N(K)}\mathsf{B}^{K'}\mathsf{P}\hat{\pi}^\mu_{s-1} - \bigotimes_{K'\in N(K)}\mathsf{B}^{K'}\mathsf{S}^N\mathsf{P}\hat{\pi}^\mu_{s-1}\|\!|\!|.\end{aligned}$$

By Corollary A.19 (applied conditionally given $\hat{\pi}^\mu_{s-1}$), we obtain

$$\sqrt{\mathbf{E}\,\|\mathsf{B}^K\mathsf{P}\tilde{\mathsf{F}}_s\hat{\pi}^\mu_{s-1} - \mathsf{B}^K\mathsf{P}\hat{\mathsf{F}}_s\hat{\pi}^\mu_{s-1}\|^2} \le \frac{8\Delta_\mathcal{K}\,\varepsilon^{-2|\mathcal{K}|_\infty}\kappa^{-2|\mathcal{K}|_\infty\Delta_\mathcal{K}}}{\sqrt{N}}.$$

The result follows immediately. □

We finally put everything together.

Theorem A.21 (Variance term)**.** *Suppose there exist $\varepsilon,\kappa>0$ with*

$$\varepsilon \le p^v(x,z^v) \le \varepsilon^{-1}, \quad \kappa \le g^v(x^v,y^v) \le \kappa^{-1} \quad \forall\, v\in V,\ x,z\in\mathbb{X},\ y\in\mathbb{Y}$$

such that

$$\varepsilon > \varepsilon_0 = \left(1 - \frac{1}{6\Delta_\mathcal{K}\Delta^2}\right)^{1/2\Delta}.$$

Let $\beta = -\log 6\Delta_\mathcal{K}\Delta^2(1-\varepsilon^{2\Delta}) > 0$. Then

$$\|\!|\!|\tilde{\pi}^x_n - \hat{\pi}^x_n\|\!|\!|_J \le \operatorname{card} J\,\frac{64\Delta_\mathcal{K}\,e^\beta}{1-e^{-\beta}}\,\frac{\varepsilon^{-4|\mathcal{K}|_\infty}\kappa^{-4|\mathcal{K}|_\infty\Delta_\mathcal{K}}}{\sqrt{N}}$$

for every $n\ge 0$, $x\in\mathbb{X}$, $K\in\mathcal{K}$ and $J\subseteq K$.

Proof. We begin with the elementary error decomposition

$$|||\tilde{\pi}_n^x - \hat{\pi}_n^x|||_J \leq \sum_{s=1}^{n} |||\tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_{s+1}\tilde{\mathsf{F}}_s\hat{\pi}_{s-1}^x - \tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_{s+1}\hat{\mathsf{F}}_s\hat{\pi}_{s-1}^x|||_J.$$

The term $s = n$ in this sum is bounded in Lemma A.17:

$$|||\tilde{\mathsf{F}}_n\hat{\pi}_{n-1}^x - \hat{\mathsf{F}}_n\hat{\pi}_{n-1}^x|||_J \leq \frac{2\kappa^{-2|\mathcal{K}|_\infty}}{\sqrt{N}}.$$

The term $s = n - 1$ is bounded in Proposition A.20:

$$|||\tilde{\mathsf{F}}_n\tilde{\mathsf{F}}_{n-1}\hat{\pi}_{s-1}^x - \tilde{\mathsf{F}}_n\hat{\mathsf{F}}_{n-1}\hat{\pi}_{s-1}^x|||_J \leq \frac{16\Delta_{\mathcal{K}}\,\varepsilon^{-2|\mathcal{K}|_\infty}\kappa^{-4|\mathcal{K}|_\infty\Delta_{\mathcal{K}}}}{\sqrt{N}}.$$

Now suppose $s < n - 1$. Then we can estimate using Corollary A.16

$$\begin{aligned}&|||\tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_{s+1}\tilde{\mathsf{F}}_s\hat{\pi}_{s-1}^x - \tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_{s+1}\hat{\mathsf{F}}_s\hat{\pi}_{s-1}^x|||_J \\ &\quad\leq \frac{4}{\varepsilon^{2|\mathcal{K}|_\infty}}\ \mathrm{card}\, J\, e^{-\beta(n-s-1)}\ \max_{K\in\mathcal{K}} \sqrt{\mathbf{E}\,\|\tilde{\mathsf{F}}_{s+1}\tilde{\mathsf{F}}_s\hat{\pi}_{s-1}^x - \tilde{\mathsf{F}}_{s+1}\hat{\mathsf{F}}_s\hat{\pi}_{s-1}^x\|_K^2}.\end{aligned}$$

Applying Proposition A.20 yields

$$\begin{aligned}&|||\tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_{s+1}\tilde{\mathsf{F}}_s\hat{\pi}_{s-1}^x - \tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_{s+1}\hat{\mathsf{F}}_s\hat{\pi}_{s-1}^x|||_J \\ &\quad\leq \mathrm{card}\, J\, e^{-\beta(n-s-1)}\ \frac{64\Delta_{\mathcal{K}}\,\varepsilon^{-4|\mathcal{K}|_\infty}\kappa^{-4|\mathcal{K}|_\infty\Delta_{\mathcal{K}}}}{\sqrt{N}}.\end{aligned}$$

Substituting the above three cases into the error decomposition and summing the geometric series yields the statement of the Theorem. □

Theorems A.12 and A.21 now immediately yield Theorem 4.2.

Appendix B

Localized Gibbs sampler particle filter: proofs

The goal of this section is to prove Theorem 5.4. What follows directly builds on the discussion presented in Section 5.6.

The key idea to bound $\|\mathsf{F}_n\rho - \tilde{\mathsf{F}}_n\rho\|_J$ is to use the one-sided Dobrushin comparison theorem (Theorem 2.12) to capture the one-sidedness that is embedded in the Gibbs samplers $\mathsf{F}_n\rho$ and $\tilde{\mathsf{F}}_n\rho$. To this end, we need to bound the one-sided coefficients C_{ij}'s and b_j's. This will be achieved, respectively, using Proposition B.1 and Proposition B.2 below; the proofs of these two propositions are based on the original Dobrushin comparison theorem (Theorem 2.11).

B.1 Preliminary steps with Dobrushin comparison theorem

The following proposition will be used to bound the C_{ij}'s coefficients in the one-sided comparison theorem.

Proposition B.1. *Suppose there exists $\varepsilon > 0$ such that*

$$\varepsilon \le p^v(x, z^v) \le \varepsilon^{-1} \qquad \text{for all } v \in V,\ x, z \in \mathbb{X}.$$

Let ν be a probability measure on $\mathbb{X}$, and suppose that

$$\mathrm{Corr}(\nu, \beta) + (1 - \varepsilon^2)e^{\beta(r+1)}\Delta \le c < 1$$

for a sufficiently small constant $\beta > 0$. Fix $n \ge 1$, and write η^v for $\eta^v_{n,\nu}$. For each $v, v' \in V$ define

$$R_{vv'} := \frac{1}{2} \sup_{\substack{x,\tilde{x}\in\mathbb{X}: \\ x^{V\setminus\{v'\}} = \tilde{x}^{V\setminus\{v'\}}}} \|\eta^v_x - \eta^v_{\tilde{x}}\|.$$

Then,

$$\max_{v\in V}\sum_{v'\in V} e^{\beta d(v,v')}R_{vv'} \le \frac{(1-\varepsilon^2)\,\Delta\, e^{\beta(r-1)}}{1-c}.$$

Proof. Henceforth, fix $n \ge 1$, $v, v' \in V$ such that $v \ne v'$ and $x, \tilde{x} \in \mathbb{X}$ such that $x^{V\setminus\{v'\}} = \tilde{x}^{V\setminus\{v'\}}$. For simplicity, write η^v for $\eta^v_{n,\nu}$. Define $I = (\{0\}\times V)\cup(1, v)$ and $\mathbb{S} = \mathbb{X}\times\mathbb{X}^v$, and the probability measures on $\mathbb{S}$

$$\rho(A) := \frac{\int \mathbf{1}_A(z,\omega)\, g^v(\omega, Y^v_n)\, \prod_{w\in V\setminus\{v\}} p^w(z,x^w)\, p^v(z,\omega)\, \nu(dz)\, \psi^v(d\omega)}{\int g^v(\omega, Y^v_n)\, \prod_{w\in V\setminus\{v\}} p^w(z,x^w)\, p^v(z,\omega)\, \nu(dz)\, \psi^v(d\omega)},$$
$$\tilde{\rho}(A) := \frac{\int \mathbf{1}_A(z,\omega)\, g^v(\omega, Y^v_n)\, \prod_{w\in V\setminus\{v\}} p^w(z,\tilde{x}^w)\, p^v(z,\omega)\, \nu(dz)\, \psi^v(d\omega)}{\int g^v(\omega, Y^v_n)\, \prod_{w\in V\setminus\{v\}} p^w(z,\tilde{x}^w)\, p^v(z,\omega)\, \nu(dz)\, \psi^v(d\omega)}.$$

By construction, for any bounded measurable function f on $\mathbb{X}^v$ we have

$$|\eta^v_x f - \eta^v_{\tilde{x}} f| = \left|\int \rho(dz, d\omega) f(\omega) - \int \tilde{\rho}(dz, d\omega) f(\omega)\right|,$$

and we will now proceed by applying the Dobrushin comparison theorem (Theorem 2.11) to bound this quantity. To this end, we must bound C_{ij} and b_i with $i = (k'', v'')$ and $j = (k''', v''')$. We distinguish two cases.

Case $k'' = 0$. In this case we have

$$\rho^i_{(z,\omega)}(A) = \frac{\int \mathbf{1}_A(z^{v''})\, \prod_{w\in N(v'')\setminus\{v\}} p^w(z,x^w)\, p^v(z,\omega)\, \nu^{v''}_z(dz^{v''})}{\int \prod_{w\in N(v'')\setminus\{v\}} p^w(z,x^w)\, p^v(z,\omega)\, \nu^{v''}_z(dz^{v''})},$$
$$\tilde{\rho}^i_{(z,\omega)}(A) = \frac{\int \mathbf{1}_A(z^{v''})\, \prod_{w\in N(v'')\setminus\{v\}} p^w(z,\tilde{x}^w)\, p^v(z,\omega)\, \nu^{v''}_z(dz^{v''})}{\int \prod_{w\in N(v'')\setminus\{v\}} p^w(z,\tilde{x}^w)\, p^v(z,\omega)\, \nu^{v''}_z(dz^{v''})}.$$

In particular, $\rho^i_{(z,x^v)} = \nu^{v''}_{z,x}$, so $C_{ij} \le C^{\nu}_{v''v'''}$ if $k''' = 0$. Moreover, as

$$\rho^i_{(z,x^v)}(A) \ge \varepsilon^2 \frac{\int \mathbf{1}_A(z^{v''}) \prod_{w\in N(v'')\setminus\{v\}} p^w(z,x^w)\, \nu^{v''}_z(dz^{v''})}{\int \prod_{w\in N(v'')\setminus\{v\}} p^w(z,x^w)\, \nu^{v''}_z(dz^{v''})},$$

we have $C_{ij} \le 1-\varepsilon^2$ if $k''' = 1$ (so $v''' = v$) and $v \in N(v'')$ by Lemma A.1, and $C_{ij} = 0$ otherwise. We therefore immediately obtain the estimate

$$\sum_{(k''',v''')\in I} e^{\beta k'''} e^{\beta d(v'',v''')} C_{(0,v'')(k''',v''')} \le \mathrm{Corr}(\nu,\beta) + (1-\varepsilon^2)e^{\beta(r+1)}.$$

On the other hand, note that $\rho^i_{(z,\omega)} = \tilde{\rho}^i_{(z,\omega)}$ if $v'' \notin N(v')$, and that we have $\rho^i_{(z,\omega)}(A) \ge \varepsilon^2\chi(A)$ and $\tilde{\rho}^i_{(z,\omega)}(A) \ge \varepsilon^2\chi(A)$, where

$$\chi(A) := \frac{\int \mathbf{1}_A(z^{v''})\, \prod_{w\in N(v'')\setminus\{v,v'\}} p^w(z,x^w)\, p^v(z,\omega)\, \nu^{v''}_z(dz^{v''})}{\int \prod_{w\in N(v'')\setminus\{v,v'\}} p^w(z,x^w)\, p^v(z,\omega)\, \nu^{v''}_z(dz^{v''})}.$$

Therefore, by Lemma A.1

$$b_i = \sup_{(z,\omega)\in\mathbb{S}} \|\rho^i_{(z,\omega)} - \tilde{\rho}^i_{(z,\omega)}\| \le \begin{cases} 0 & \text{for } v'' \notin N(v'), \\ 2(1-\varepsilon^2) & \text{otherwise.} \end{cases}$$

Case $k'' = 1$. In this case we have

$$\rho^i_{(z,\omega)}(A) = \tilde{\rho}^i_{(z,\omega)}(A) = \frac{\int \mathbf{1}_A(\omega)\, g^v(\omega, Y^v_n)\, p^v(z,\omega)\, \psi^v(d\omega)}{\int g^v(\omega, Y^v_n)\, p^v(z,\omega)\, \psi^v(d\omega)}.$$

Thus $b_i = 0$, and estimating as above we obtain $C_{ij} \le 1-\varepsilon^2$ whenever $k''' = 0$ and $v''' \in N(v)$, and $C_{ij} = 0$ otherwise. In particular, we obtain

$$\sum_{(k''',v''')\in I} e^{\beta|1-k'''|} e^{\beta d(v,v''')} C_{(1,v)(k''',v''')} \le (1-\varepsilon^2) e^{\beta(r+1)} \Delta.$$

Combining the above two cases and the assumption of the Proposition yields

$$\max_{(k'',v'')\in I} \sum_{(k''',v''')\in I} e^{\beta\{|k''-k'''|+d(v'',v''')\}} C_{(k'',v'')(k''',v''')} \le c.$$

Applying Theorem 2.11 gives

$$\begin{aligned} \|\eta^v_x - \eta^v_{\tilde{x}}\| &= \sup_{f\in\mathfrak{X}^v : |f|\le 1} \left| \int \rho(dz, d\omega) f(\omega) - \int \tilde{\rho}(dz, d\omega) f(\omega) \right| \\ &\le 2(1-\varepsilon^2) \sum_{v''\in N(v')} D_{(1,v)(0,v'')}. \end{aligned}$$

As the previous bound does not depend on the choice of $x, \tilde{x} \in \mathbb{X}$, as long as $x^{V\setminus\{v'\}} = \tilde{x}^{V\setminus\{v'\}}$, we have

$$R_{vv'} = \frac{1}{2} \sup_{\substack{x,\tilde{x}\in\mathbb{X}: \\ x^{V\setminus\{v'\}} = \tilde{x}^{V\setminus\{v'\}}}} \|\eta^v_x - \eta^v_{\tilde{x}}\| \le (1-\varepsilon^2) \sum_{v''\in N(v')} D_{(1,v)(0,v'')}.$$

By Lemma 2.13 it follows that

$$\begin{aligned} \max_{v\in V} \sum_{v'\in V} e^{\beta d(v,v')} R_{vv'} &\le (1-\varepsilon^2) \max_{v\in V} \sum_{v'\in V} e^{\beta d(v,v')} \sum_{v''\in N(v')} D_{(1,v)(0,v'')} \\ &\le (1-\varepsilon^2) \max_{v\in V} \sum_{v''\in V} e^{\beta d(v,v'')+\beta r} D_{(1,v)(0,v'')} \sum_{v'\in V} \mathbf{1}_{v''\in N(v')} \\ &\le (1-\varepsilon^2)\, \Delta\, e^{\beta(r-1)} \max_{v\in V} \sum_{v''\in V} e^{\beta d(v,v'')+\beta} D_{(1,v)(0,v'')} \\ &\le \frac{(1-\varepsilon^2)\, \Delta\, e^{\beta(r-1)}}{1-c}, \end{aligned}$$

and, in particular,

$$\max_{v\in V} \sum_{v'\in V} R_{vv'} \le \frac{(1-\varepsilon^2)\, \Delta\, e^{\beta(r-1)}}{1-c}.$$

□

The following proposition will be used to bound the b_j's coefficients in the one-sided comparison theorem.

Proposition B.2. *Suppose there exists $\varepsilon > 0$ such that*

$$\varepsilon \le p^v(x, z^v) \le \varepsilon^{-1} \qquad \text{for all } v \in V,\ x, z \in \mathbb{X}.$$

Let ν be a probability measure on $\mathbb{X}$, and suppose that

$$\mathrm{Corr}(\nu, \beta) + (1 - \varepsilon^2) e^{\beta(r+1)} \Delta \le c < 1$$

for a sufficiently small constant $\beta > 0$. Fix $n \ge 1$, and write η^v for $\eta^v_{n,\nu}$. Then, for each $v \in V$ we have

$$\sup_{x \in \mathbb{X}} \|\eta^v_x - \tilde{\eta}^v_x\| \le 2 \, \frac{(1 - \varepsilon^{2\Delta}) e^{-\beta(2-r)}}{1 - c} \, e^{-\beta b}.$$

Proof. Henceforth, fix $n \ge 1$, $v \in V$, $x \in \mathbb{X}$. For simplicity, write η^v for $\eta^v_{n,\nu}$ and $\tilde{\eta}^v$ for $\tilde{\eta}^v_{n,\nu}$. Define $I = (\{0\} \times V) \cup (1, v)$ and $\mathbb{S} = \mathbb{X} \times \mathbb{X}^v$, and the probability measures on $\mathbb{S}$

$$\rho(A) := \frac{\int \mathbf{1}_A(z, \omega)\, g^v(\omega, Y^v_n) \prod_{w \in V \setminus \{v\}} p^w(z, x^w)\, p^v(z, \omega)\, \nu(dz)\, \psi^v(d\omega)}{\int g^v(\omega, Y^v_n) \prod_{w \in V \setminus \{v\}} p^w(z, x^w)\, p^v(z, \omega)\, \nu(dz)\, \psi^v(d\omega)},$$
$$\tilde{\rho}(A) := \frac{\int \mathbf{1}_A(z, \omega)\, g^v(\omega, Y^v_n) \prod_{w \in N_b(v) \setminus \{v\}} p^w(z, x^w)\, p^v(z, \omega)\, \nu(dz)\, \psi^v(d\omega)}{\int g^v(\omega, Y^v_n) \prod_{w \in N_b(v) \setminus \{v\}} p^w(z, x^w)\, p^v(z, \omega)\, \nu(dz)\, \psi^v(d\omega)}.$$

By construction, for any bounded measurable function f on $\mathbb{X}^v$ we have

$$|\eta^v_x f - \tilde{\eta}^v_x f| = \left| \int \rho(dz, d\omega) f(\omega) - \int \tilde{\rho}(dz, d\omega) f(\omega) \right|,$$

and we now proceed by applying the Dobrushin comparison theorem (Theorem 2.11) to bound this quantity. To this end, we must bound C_{ij} and b_i with $i = (k', v')$ and $j = (k'', v'')$. We distinguish two cases.

Case $k' = 0$. In this case we have

$$\rho^i_{(z, x^v)}(A) = \frac{\int \mathbf{1}_A(z^{v'}) \prod_{w \in N(v')} p^w(z, x^w)\, \nu^{v'}_z(dz^{v'})}{\int \prod_{w \in N(v')} p^w(z, x^w)\, \nu^{v'}_z(dz^{v'})},$$
$$\tilde{\rho}^i_{(z, x^v)}(A) = \frac{\int \mathbf{1}_A(z^{v'}) \prod_{w \in N(v') \cap N_b(v)} p^w(z, x^w)\, \nu^{v'}_z(dz^{v'})}{\int \prod_{w \in N(v') \cap N_b(v)} p^w(z, x^w)\, \nu^{v'}_z(dz^{v'})}.$$

In particular, $\rho^i_{(z, x^v)} = \nu^{v'}_{z,x}$, so $C_{ij} \le C^\nu_{v'v''}$ if $k'' = 0$. Moreover, as

$$\rho^i_{(z, x^v)}(A) \ge \varepsilon^2 \, \frac{\int \mathbf{1}_A(z^{v'}) \prod_{w \in N(v') \setminus \{v\}} p^w(z, x^w)\, \nu^{v'}_z(dz^{v'})}{\int \prod_{w \in N(v') \setminus \{v\}} p^w(z, x^w)\, \nu^{v'}_z(dz^{v'})},$$

we have $C_{ij} \leq 1-\varepsilon^2$ if $k''=1$ (so $v''=v$) and $v \in N(v')$ by Lemma A.1, and $C_{ij}=0$ otherwise. We therefore immediately obtain the estimate

$$\sum_{(k'',v'')\in I} e^{\beta k''} e^{\beta d(v',v'')} C_{(0,v')(k'',v'')} \leq \mathrm{Corr}(\nu,\beta) + (1-\varepsilon^2) e^{\beta(r+1)}.$$

On the other hand, note that $\rho^i_{(z,x^v)} = \tilde{\rho}^i_{(z,x^v)}$ if $N(v') \subseteq N_b(v)$, and that we have $\rho^i_{(z,x^v)} \geq \varepsilon^{2\Delta} \nu_z^{v'}$ and $\tilde{\rho}^i_{(z,x^v)} \geq \varepsilon^{2\Delta} \nu_z^{v'}$. Therefore, by Lemma A.1

$$b_i = \sup_{(z,x^v)\in\mathbb{S}} \|\rho^i_{(z,x^v)} - \tilde{\rho}^i_{(z,x^v)}\| \leq \begin{cases} 0 & \text{for } v' \in N_b(v) \backslash \partial N_b(v), \\ 2(1-\varepsilon^{2\Delta}) & \text{otherwise.} \end{cases}$$

Case $k'=1$. In this case we have

$$\rho^i_{(z,x^v)}(A) = \tilde{\rho}^i_{(z,x^v)}(A) = \frac{\int \mathbf{1}_A(x^v)\, g^v(x^v, Y_n^v)\, p^v(z,x^v)\, \psi^v(dx^v)}{\int g^v(x^v, Y_n^v)\, p^v(z,x^v)\, \psi^v(dx^v)}.$$

Thus $b_i = 0$, and estimating as above we obtain $C_{ij} \leq 1-\varepsilon^2$ whenever $k''=0$ and $v'' \in N(v)$, and $C_{ij}=0$ otherwise. In particular, we obtain

$$\sum_{(k'',v'')\in I} e^{\beta|1-k''|} e^{\beta d(v,v'')} C_{(1,v)(k'',v'')} \leq (1-\varepsilon^2) e^{\beta(r+1)} \Delta.$$

Combining the above two cases and the assumption of the Proposition yields

$$\max_{(k',v')\in I} \sum_{(k'',v'')\in I} e^{\beta\{|k'-k''|+d(v',v'')\}} C_{(k',v')(k'',v'')} \leq c.$$

Applying Theorem 2.11 and Lemma 2.13 gives

$$\begin{aligned} \|\eta_x^v - \tilde{\eta}_x^v\| &= \sup_{f\in\mathfrak{X}^v : |f|\leq 1} \left| \int \rho(dz, d\omega) f(\omega) - \int \tilde{\rho}(dz, d\omega) f(\omega) \right| \\ &\leq 2(1-\varepsilon^{2\Delta}) \sum_{v'\in V\backslash(N_b(v)\backslash\partial N_b(v))} D_{(1,v)(0,v')} \\ &\leq \frac{2}{1-c} (1-\varepsilon^{2\Delta}) e^{-\beta}\, e^{-\beta d(v,\partial N_b(v))} \\ &\leq \frac{2}{1-c} (1-\varepsilon^{2\Delta}) e^{-\beta(2-r)}\, e^{-\beta b}, \end{aligned}$$

where in the last inequality we used the fact that $d(v, \partial N_b(v)) \geq b-r+1$. As the choice of $x \in \mathbb{X}$ was arbitrary, we get

$$\sup_{x\in\mathbb{X}} \|\eta_x^v - \tilde{\eta}_x^v\| \leq 2\, \frac{(1-\varepsilon^{2\Delta}) e^{-\beta(2-r)}}{1-c}\, e^{-\beta b}.$$

□

B.2 Proof of Theorem 5.4 with one-sided Dobrushin comparison theorem

Using Proposition B.1 and Proposition B.2 we can now apply the one-sided Dobrushin comparison theorem (Theorem 2.12) to analyze the quantity $\|\mathsf{F}_n\nu - \tilde{\mathsf{F}}_n\nu\|_J$ and to provide a bound that is spatially homogeneous in $J \subseteq V$. As explained in Section 5.6, the key intuition behind the following proof is that both the filter recursion and the approximate filter recursion can be phrased in terms of Gibbs samplers, which can then be easily compared.

Theorem B.3. *Suppose there exists $\varepsilon > 0$ such that*

$$\varepsilon \le p^v(x, z^v) \le \varepsilon^{-1} \qquad \text{for all } v \in V,\ x, z \in \mathbb{X}.$$

Let ν be a probability measure on $\mathbb{X}$, and suppose that

$$\begin{aligned} &\mathrm{Corr}(\nu, \beta) + (1-\varepsilon^2)e^{\beta(r+1)}\Delta \le c < 1, \\ &\frac{(1-\varepsilon^2)\, e^{\beta(r+1)}\, \Delta}{1-c} \le c' < 1, \end{aligned}$$

for a sufficiently small constant $\beta > 0$. Then, for each $n \ge 1$ and $J \subseteq V$ we have

$$\|\mathsf{F}_n\nu - \tilde{\mathsf{F}}_n\nu\|_J \le 2\ \mathrm{card}\, J \left(\frac{e^{-\beta(2-r)}(1-\varepsilon^{2\Delta})}{(1-c)(1-c')}\, e^{-\beta b} + \frac{1}{1-c'}\, c'^m \right).$$

Proof. Fix $n \ge 1$ and $J \subseteq V$. To lighten the notation, we write η^v for $\eta^v_{n,\nu}$ and G^v for $G^v_{n,\nu}$, and analogously for $\tilde{\eta}^v$ and $\tilde{G}^v$. By construction, for each $v \in V$ the kernel $G^v_{n,\rho}$ leaves the measure $\mathsf{F}_n\rho$ invariant, that is,

$$(\mathsf{F}_n\rho)G^v_{n,\rho} = \mathsf{F}_n\rho.$$

Hence, we can express the filter recursion as m sweeps of a Gibbs sampler, namely,

$$\mathsf{F}_n\rho = (\mathsf{F}_n\rho)(G^{v_1}_{n,\rho} \cdots G^{v_d}_{n,\rho})^m.$$

On the other hand, the approximate Gibbs sampler filter recursion reads

$$\tilde{\mathsf{F}}_n\rho := \rho\,(\tilde{G}^{v_1}_{n,\rho} \cdots \tilde{G}^{v_d}_{n,\rho})^m.$$

Therefore, we can decompose the one-step error between filter and approximate filter as

$$\begin{aligned} \|(\mathsf{F}_n\nu) - (\tilde{\mathsf{F}}_n\nu)\|_J &= \sup_{f \in \mathcal{X}^J : |f| \le 1} |(\mathsf{F}_n\nu)(G^{v_1} \cdots G^{v_d})^m f - \nu(\tilde{G}^{v_1} \cdots \tilde{G}^{v_d})^m f| \\ &\le \sup_{f \in \mathcal{X}^J : |f| \le 1} \sup_{z, \tilde{z} \in \mathbb{X}} |\delta_z(G^{v_1} \cdots G^{v_d})^m f - \delta_{\tilde{z}}(\tilde{G}^{v_1} \cdots \tilde{G}^{v_d})^m f| \\ &\le \sup_{z \in \mathbb{X}} \|\delta_z(G^{v_1} \cdots G^{v_d})^m - \delta_z(\tilde{G}^{v_1} \cdots \tilde{G}^{v_d})^m\|_J \\ &\quad + \sup_{z, \tilde{z} \in \mathbb{X}} \|\delta_z(G^{v_1} \cdots G^{v_d})^m - \delta_{\tilde{z}}(G^{v_1} \cdots G^{v_d})^m\|_J. \end{aligned} \tag{B.1}$$

We first analyze the first term on the right side of (B.1), which will give us a bound depending on the approximation parameter b. Henceforth, fix $z \in \mathbb{X}$. For any bounded measurable function f on $\mathbb{X}$ we have

$$\delta_z(G^{v_1}\cdots G^{v_d})^m f = \int \delta_z(dx_0) \prod_{\ell=1}^{m}\prod_{k=1}^{d} \eta^{v_k}(x_\ell^{\{v_1,\dots,v_{k-1}\}} x_{\ell-1}^{\{v_k,\dots,v_d\}}, dx_\ell^{v_k})\, f(x_m),$$
$$\delta_z(\tilde{G}^{v_1}\cdots \tilde{G}^{v_d})^m f = \int \delta_z(dx_0) \prod_{\ell=1}^{m}\prod_{k=1}^{d} \tilde{\eta}^{v_k}(x_\ell^{\{v_1,\dots,v_{k-1}\}} x_{\ell-1}^{\{v_k,\dots,v_d\}}, dx_\ell^{v_k})\, f(x_m).$$

Define $I := \bigcup_{\ell=0}^{m}(\{\ell\}\times V)$ and $\mathbb{S} := \bigotimes_{\ell=0}^{m}\mathbb{X}$. Define the probability measures on $\mathbb{S}$

$$\rho(A) := \int \delta_z(dx_0) \prod_{\ell=1}^{m}\prod_{k=1}^{d} \eta^{v_k}(x_\ell^{\{v_1,\dots,v_{k-1}\}} x_{\ell-1}^{\{v_k,\dots,v_d\}}, dx_\ell^{v_k})\, \mathbf{1}_A(x),$$
$$\tilde{\rho}(A) := \int \delta_z(dx_0) \prod_{\ell=1}^{m}\prod_{k=1}^{d} \tilde{\eta}^{v_k}(x_\ell^{\{v_1,\dots,v_{k-1}\}} x_{\ell-1}^{\{v_k,\dots,v_d\}}, dx_\ell^{v_k})\, \mathbf{1}_A(x).$$

By construction, we have

$$|\delta_z(G^{v_1}\cdots G^{v_d})^m f - \delta_z(\tilde{G}^{v_1}\cdots \tilde{G}^{v_d})^m f| = \left|\int \rho(dx) f(x_m) - \int \tilde{\rho}(dx) f(x_m)\right|.$$

We want to use Theorem 2.12 to bound this quantity. To this end, let τ be defined as

$$\tau : i = (\ell, v_k) \in I \longrightarrow \tau(i) = \ell d + k,$$

and for each $i \in I$, $x \in \mathbb{S}$, let

$$\gamma_x^i(A) := \rho(X^i \in A | X^{I_{\le\tau(i)}\setminus\{i\}} = x^{I_{\le\tau(i)}\setminus\{i\}}),$$
$$\tilde{\gamma}_x^i(A) := \tilde{\rho}(X^i \in A | X^{I_{\le\tau(i)}\setminus\{i\}} = x^{I_{\le\tau(i)}\setminus\{i\}}).$$

We immediately find that for each $x \in \mathbb{S}$, $\ell \in \{1,\dots,m\}$, $k \in \{1,\dots,d\}$, we have

$$\gamma_x^{(0,v)}(A) = \tilde{\gamma}_x^{(0,v)}(A) = \delta_{z^v}(A),$$
$$\gamma_x^{(\ell,v_k)}(A) = \eta^{v_k}(x_\ell^{\{v_1,\dots,v_{k-1}\}} x_{\ell-1}^{\{v_k,\dots,v_d\}}, A),$$
$$\tilde{\gamma}_x^{(\ell,v_k)}(A) = \tilde{\eta}^{v_k}(x_\ell^{\{v_1,\dots,v_{k-1}\}} x_{\ell-1}^{\{v_k,\dots,v_d\}}, A).$$

Recall the following definition from Proposition B.1:

$$R_{vv'} := \frac{1}{2} \sup_{\substack{x,\tilde{x}\in\mathbb{X}:\\ x^{I\setminus\{v'\}}=\tilde{x}^{I\setminus\{v'\}}}} \|\eta_x^v - \eta_{\tilde{x}}^v\| \qquad \text{for } v, v' \in V.$$

It is easy to check that for each $i, j \in I$ we have

$$C_{ij} = \begin{cases} R_{vv'} & \text{if } i = (\ell, v), j = (\ell', v') \text{ for } 0 \le \tau(i) - \tau(j) \le d-1, \\ 0 & \text{otherwise,} \end{cases}$$

and for each $j \in I$ we have

$$b_j = \begin{cases} \sup_{x\in\mathbb{X}} \|\eta_x^v - \tilde\eta_x^v\| & \text{if } j = (\ell, v),\ \ell \geq 1, \\ 0 & \text{otherwise.} \end{cases}$$

So, by Theorem 2.12 and Proposition B.2 (to bound the b_j's) we have

$$\begin{aligned} \|\delta_z(G^{v_1}\cdots G^{v_d})^m - \delta_z(\tilde G^{v_1}\cdots \tilde G^{v_d})^m\|_J &\leq \sum_{v\in J}\sum_{j\in I} D_{(m,v)j}\, b_j \\ &\leq 2\,\frac{(1-\varepsilon^{2\Delta})e^{-\beta(2-r)}}{1-c}\, e^{-\beta b}\sum_{v\in J}\sum_{j\in I} D_{(m,v)j}. \end{aligned}$$

Moreover, by Proposition B.1 we have

$$\max_{i\in I}\sum_{j\in I} C_{ij} = \max_{v\in V}\sum_{v'\in V} R_{vv'} \leq \frac{(1-\varepsilon^2)\,\Delta\, e^{\beta(r-1)}}{1-c} \leq c' < 1,$$

from which by Lemma 2.13 it follows that

$$\max_{i\in I}\sum_{j\in I} D_{ij} \leq \frac{1}{1-c'}.$$

We finally obtain

$$\|\delta_z(G^{v_1}\cdots G^{v_d})^m - \delta_z(\tilde G^{v_1}\cdots \tilde G^{v_d})^m\|_J \leq 2\ \text{card}\, J\ \frac{(1-\varepsilon^{2\Delta})e^{-\beta(2-r)}}{(1-c)(1-c')}\, e^{-\beta b}. \tag{B.2}$$

We now analyze the second term on the right side of (B.1), which will give us a bound depending on the iteration step m. Henceforth, fix $z, \tilde z \in \mathbb{X}$. Define the probability measures on $\mathbb{S}$

$$\begin{aligned} \rho(A) &:= \int \delta_z(dx_0)\prod_{\ell=1}^m\prod_{k=1}^d \eta^{v_k}(x_\ell^{\{v_1,\dots,v_{k-1}\}} x_{\ell-1}^{\{v_k,\dots,v_d\}}, dx_\ell^{v_k})\,\mathbf{1}_A(x), \\ \tilde\rho(A) &:= \int \delta_{\tilde z}(dx_0)\prod_{\ell=1}^m\prod_{k=1}^d \eta^{v_k}(x_\ell^{\{v_1,\dots,v_{k-1}\}} x_{\ell-1}^{\{v_k,\dots,v_d\}}, dx_\ell^{v_k})\,\mathbf{1}_A(x). \end{aligned}$$

By construction we have, for any bounded measurable function f on $\mathbb{X}$,

$$|\delta_z(G^{v_1}\cdots G^{v_d})^m f - \delta_{\tilde z}(G^{v_1}\cdots G^{v_d})^m f| = \left|\int \rho(dx) f(x_m) - \int \tilde\rho(dx) f(x_m)\right|.$$

In the present case we find the following expressions for the one-sided conditional distributions, for each $x \in \mathbb{S}$, $\ell \in \{1,\dots,m\}$, $k \in \{1,\dots,d\}$,

$$\begin{aligned} \gamma_x^{(0,v)}(A) &= \delta_{z^v}(A), \\ \tilde\gamma_x^{(0,v)}(A) &= \delta_{\tilde z^v}(A), \\ \gamma_x^{(\ell,v_k)}(A) &= \tilde\gamma_x^{(\ell,v_k)}(A) = \eta^{v_k}(x_\ell^{\{v_1,\dots,v_{k-1}\}} x_{\ell-1}^{\{v_k,\dots,v_d\}}, A). \end{aligned}$$

As before, for each $i, j \in I$ we have

$$C_{ij} = \begin{cases} R_{vv'} & \text{if } i = (\ell, v), j = (\ell', v') \text{ for } 0 \le \tau(i) - \tau(j) \le d-1, \\ 0 & \text{otherwise,} \end{cases}$$

and for each $j \in I$ we now have

$$b_j \le \begin{cases} 2 & \text{if } \tau(j) = 0, \\ 0 & \text{otherwise.} \end{cases}$$

By Theorem 2.12 we find

$$\|\delta_z (G^{v_1} \cdots G^{v_d})^m - \delta_{\tilde{z}} (G^{v_1} \cdots G^{v_d})^m\|_J \le \sum_{v \in J} \sum_{j \in I} D_{(m,v)j}\, b_j \le 2 \sum_{v \in J} \sum_{v' \in V} D_{(m,v)(0,v')}.$$

Proceeding as above, by Proposition B.1 we have

$$\max_{i \in I} \sum_{j \in I} C_{ij} \le c' < 1,$$

from which it follows that

$$\max_{v \in V} \sum_{v' \in V} D_{(m,v)(0,v')} = \sum_{n=m}^{\infty} C^n_{(m,v)(0,v')} \le \frac{c'^m}{1-c'},$$

where we have used that $C_{ij} \neq 0$ only if $0 \le \tau(i) - \tau(j) \le d-1$. We finally obtain

$$\|\delta_z (G^{v_1} \cdots G^{v_d})^m - \delta_{\tilde{z}} (G^{v_1} \cdots G^{v_d})^m\|_J = 2 \operatorname{card} J \, \frac{c'^m}{1-c'}. \tag{B.3}$$

As the choice of $z, \tilde{z} \in \mathbb{X}$ is arbitrary, together (B.2) and (B.3) yield the statement of the Theorem. □

The proof of Theorem 5.4 follows as an immediate consequence of Theorem B.3.

Proof of Theorem 5.4. In Theorem B.3, choose $c = \frac{1}{2}$ and $c' = \frac{1}{4}$. Let

$$\frac{(1-\varepsilon^2)\, e^{\beta(r+1)}\, \Delta}{1-c} = c',$$

from which we get $\beta = \frac{1}{r+1} \log \frac{1}{8\Delta(1-\varepsilon^2)} > 0$, as $\varepsilon > \varepsilon_0 := \sqrt{1 - \frac{1}{8\Delta}}$. As $\operatorname{Corr}(\nu, \beta) \le \frac{1}{4}$ by assumption, we find

$$\operatorname{Corr}(\nu, \beta) + (1-\varepsilon^2) e^{\beta(r+1)} \Delta \le \frac{1}{4} + c'(1-c) = \frac{3}{8} \le \frac{1}{2} \equiv c.$$

Hence, both assumptions in Theorem B.3 hold, and for each $n \geq 1$ and $J \subseteq V$ we get

$$\begin{aligned}\|\mathsf{F}_n\nu - \tilde{\mathsf{F}}_n\nu\|_J &\leq 2 \operatorname{card} J \left(\frac{8}{3} e^{-\beta(2-r)} \left(1 - \varepsilon^{2\Delta}\right) e^{-\beta b} + \frac{4}{3} \left(\frac{1}{4}\right)^m\right) \\ &\leq \frac{\alpha}{2} \operatorname{card} J \left(e^{-\beta b} + e^{-(\log 4)m}\right) \\ &\leq \alpha \operatorname{card} J \, e^{-\gamma \min\{b,m\}},\end{aligned}$$

where

$$\begin{aligned}\alpha &:= 4 \left(\frac{8}{3} \left(8\Delta(1-\varepsilon^2)\right)^{\frac{2-r}{r+1}} \left(1 - \varepsilon^{2\Delta}\right) + \frac{4}{3}\right), \\ \gamma &:= \min\left\{\frac{1}{r+1} \log \frac{1}{8\Delta(1-\varepsilon^2)}, \log 4\right\}.\end{aligned}$$

□

Appendix C

Comparison theorems for Gibbs measures: proofs

The first part of this appendix (Sections C.1-C.5) is devoted to providing the proofs for the generalized comparison theorems introduced in Chapter 6 (Theorem 6.4, Corollary 6.8, and Theorem 6.12). The second part of this appendix (Sections C.6) is devoted to developing the application of the generalized comparison theorems to block particle filters (Theorem 6.13).

C.1 General comparison principle

The proof of Theorem 6.4 is derived from a general comparison principle for Markov chains that will be formalized in this section. The basic idea behind this principle is to consider two transition kernels G and $\tilde{G}$ on $\mathbb{S}$ such that $\rho G = G$ and $\tilde{\rho}\tilde{G} = \tilde{\rho}$. One should think of G as the transition kernel of a Markov chain that admits ρ as its invariant measure, and similarly for $\tilde{G}$. The comparison principle of this section provides a general method to bound the difference between the invariant measures ρ and $\tilde{\rho}$ in terms of the transition kernels G and $\tilde{G}$. In the following sections, we will apply this principle to a specific choice of G and $\tilde{G}$ that is derived from the coupled update rule.

We begin by introducing a standard notion in the analysis of high-dimensional Markov chains, cf. [23] (note that our indices are reversed as compared to the definition in [23]).

Definition C.1. $(V_{ij})_{i,j\in I}$ *is called a* Wasserstein matrix *for a transition kernel* G *on* $\mathbb{S}$ *if*

$$\operatorname{osc}_j Gf \le \sum_{i\in I} \operatorname{osc}_i f \, V_{ij}$$

for every $j \in I$ *and bounded and measurable quasilocal function* f.

We now state our general comparison principle.

Proposition C.2. *Let G and $\tilde{G}$ be transition kernels on $\mathbb{S}$ such that $\rho G = \rho$ and $\tilde{\rho}\tilde{G} = \tilde{\rho}$, and let Q_x be a coupling between the measures G_x and $\tilde{G}_x$ for every $x \in \mathbb{S}$. Assume that G is quasilocal, and let V be a Wasserstein matrix for G. Then we have*

$$|\rho f - \tilde{\rho} f| \le \sum_{i,j\in I} \operatorname{osc}_i f \, N^{(n)}_{ij} \int^* \tilde{\rho}(dx)\, Q_x \eta_j + \sum_{i,j\in I} \operatorname{osc}_i f \, V^n_{ij} \, (\rho \otimes \tilde{\rho})\eta_j,$$

where we defined

$$N^{(n)} := \sum_{k=0}^{n-1} V^k,$$

for any bounded and measurable quasilocal function f and $n \ge 1$.

Theorem 6.4 will be derived from this result. Roughly speaking, we will design the transition kernel G such that $V = I - W + R$ is a Wasserstein matrix; then assumption (6.1) implies that the second term in Proposition C.2 vanishes as $n \to \infty$, and the result of Theorem 6.4 reduces to some matrix algebra (as will be explained below, however, a more complicated argument is needed to obtain Theorem 6.4 in full generality).

To prove Proposition C.2 we require a simple lemma.

Lemma C.3. *Let Q be a coupling of probability measures μ, ν on $\mathbb{S}$. Then*

$$|\mu f - \nu f| \le \sum_{i\in I} \operatorname{osc}_i f \, Q\eta_i$$

for every bounded and measurable quasilocal function f.

Proof. Let $J \in \mathfrak{I}$. Enumerate its elements arbitrarily as $J = \{j_1, \ldots, j_r\}$, and define $J_k = \{j_1, \ldots, j_k\}$ for $1 \le k \le r$ and $J_0 = \varnothing$. Then we can evidently estimate

$$|f^J_x(z) - f^J_x(\tilde{z})| \le \sum_{k=1}^{r} |f^J_x(z^{J_k}\tilde{z}^{J\setminus J_k}) - f^J_x(z^{J_{k-1}}\tilde{z}^{J\setminus J_{k-1}})| \le \sum_{j\in J} \operatorname{osc}_j f \, \eta_j(z_j, \tilde{z}_j).$$

As f is quasilocal, we can let $J \uparrow I$ to obtain

$$|f(z) - f(\tilde{z})| \le \sum_{i\in I} \operatorname{osc}_i f \, \eta_i(z_i, \tilde{z}_i).$$

The result follows readily as $|\mu f - \nu f| \le \int |f(z) - f(\tilde{z})| \, Q(dz, d\tilde{z})$. □

We now proceed to the proof of Proposition C.2.

Proof of Proposition C.2. We begin by writing

$$\begin{aligned}
|\rho f - \tilde{\rho} f| &= |\rho G^n f - \tilde{\rho}\tilde{G}^n f| \\
&\le \sum_{k=0}^{n-1} |\tilde{\rho}\tilde{G}^{n-k-1} G^{k+1} f - \tilde{\rho}\tilde{G}^{n-k} G^k f| + |\rho G^n f - \tilde{\rho} G^n f| \\
&= \sum_{k=0}^{n-1} |\tilde{\rho} G G^k f - \tilde{\rho}\tilde{G} G^k f| + |\rho G^n f - \tilde{\rho} G^n f|.
\end{aligned}$$

As G is assumed quasilocal, $G^k f$ is quasilocal, and thus Lemma C.3 yields

$$\begin{aligned}|\tilde\rho G G^k f - \tilde\rho \tilde G G^k f| &\le \int \tilde\rho(dx)\,|G_x G^k f - \tilde G_x G^k f| \\ &\le \int^* \tilde\rho(dx) \sum_{j\in I} \mathrm{osc}_j G^k f\, Q_x\eta_j \\ &\le \sum_{i,j\in I} \mathrm{osc}_i f\, V^k_{ij} \int^* \tilde\rho(dx)\, Q_x\eta_j.\end{aligned}$$

Similarly, as $\rho\otimes\tilde\rho$ is a coupling of $\rho,\tilde\rho$, we obtain by Lemma C.3

$$|\rho G^n f - \tilde\rho G^n f| \le \sum_{j\in I} \mathrm{osc}_j G^n f\,(\rho\otimes\tilde\rho)\eta_j \le \sum_{i,j\in I} \mathrm{osc}_i f\, V^n_{ij}\,(\rho\otimes\tilde\rho)\eta_j.$$

Thus the proof is complete. □

C.2 Gibbs samplers

To put Proposition C.2 to good use, we must construct transition kernels G and $\tilde G$ for which ρ and $\tilde\rho$ are invariant, and that admit tractable estimates for the quantities in the comparison theorem in terms of the coupled update rule $(\gamma^J,\tilde\gamma^J,Q^J,\hat Q^J)_{J\in\mathcal{J}}$ and the weights $(w_J)_{J\in\mathcal{J}}$. To this end, we will use a standard construction called the *Gibbs sampler*: in each time step, we draw a region $J\in\mathcal{J}$ with probability $v_J\propto w_J$, and then apply the transition kernel γ^J to the current configuration. This readily defines a transition kernel G for which ρ is G-invariant (as ρ is γ^J-invariant for every $J\in\mathcal{J}$). The construction for $\tilde G$ is identical. As will be explained below, this is not the most natural construction for the proof of our main result; however, it will form the basis for further computations.

We fix throughout this section a coupled update rule $(\gamma^J,\tilde\gamma^J,Q^J,\hat Q^J)_{J\in\mathcal{J}}$ for $(\rho,\tilde\rho)$ and weights $(w_J)_{J\in\mathcal{J}}$ satisfying the assumptions of Theorem 6.4. Let $\mathbf{v}=(v_J)_{J\in\mathcal{J}}$ be a sequence of nonnegative weights such that $\sum_J v_J\le 1$. We define the Gibbs samplers

$$\begin{aligned}G^{\mathbf{v}}_x(A) &:= \Big(1-\sum_{J\in\mathcal{J}} v_J\Big)\mathbf{1}_A(x) + \sum_{J\in\mathcal{J}} v_J \int \mathbf{1}_A(z^J x^{I\setminus J})\,\gamma^J_x(dz^J),\\ \tilde G^{\mathbf{v}}_x(A) &:= \Big(1-\sum_{J\in\mathcal{J}} v_J\Big)\mathbf{1}_A(x) + \sum_{J\in\mathcal{J}} v_J \int \mathbf{1}_A(z^J x^{I\setminus J})\,\tilde\gamma^J_x(dz^J).\end{aligned}$$

Evidently $G^{\mathbf{v}}$ and $\tilde G^{\mathbf{v}}$ are transition kernels on $\mathbb{S}$, and $\rho G^{\mathbf{v}}=\rho$ and $\tilde\rho\tilde G^{\mathbf{v}}=\tilde\rho$ by construction. To apply Proposition C.2, we must establish some basic properties.

Lemma C.4. *Assume that γ^J is quasilocal for every $J\in\mathcal{J}$. Then $G^{\mathbf{v}}$ is quasilocal.*

Proof. Let $f:\mathbb{S}\to\mathbb{S}$ be a bounded and measurable quasilocal function. It evidently suffices to show that $\gamma^J f^J$ is quasilocal for every $J\in\mathcal{J}$. To this end, let us fix $J\in\mathcal{J}$, $x,z\in\mathbb{S}$, and $J_1,J_2,\ldots\in\mathcal{I}$ such that $J_1\subseteq J_2\subseteq\cdots$ and $\bigcup_i J_i = I$. Then we have

$$\gamma^J_{z^{J_i}x^{I\setminus J_i}} \xrightarrow{i\to\infty} \gamma^J_z \quad \text{setwise}$$

as γ^J is quasilocal. On the other hand, we have

$$f^J_{z^{J_i}x^{I\setminus J_i}} \xrightarrow{i\to\infty} f^J_z \quad \text{pointwise}$$

as f is quasilocal. Thus by [43, Proposition 18, p. 270] we obtain

$$\gamma^J_{z^{J_i}x^{I\setminus J_i}} f^J_{z^{J_i}x^{I\setminus J_i}} \xrightarrow{i\to\infty} \gamma^J_z f^J_z.$$

As the choice of x, z and $(J_i)_{i\geq 1}$ is arbitrary, the result follows. □

Lemma C.5. *Assume that γ^J is quasilocal for every $J \in \mathcal{J}$, and define*

$$W^{\mathbf{v}}_{ij} := \mathbf{1}_{i=j} \sum_{J\in\mathcal{J}: i\in J} v_J,$$

$$R^{\mathbf{v}}_{ij} := \sup_{\substack{x,z\in\mathbb{S}:\\ x^{I\setminus\{j\}}=z^{I\setminus\{j\}}}} \frac{1}{\eta_j(x_j,z_j)} \sum_{J\in\mathcal{J}: i\in J} v_J\, Q^J_{x,z}\eta_i.$$

Then $V^{\mathbf{v}} = I - W^{\mathbf{v}} + R^{\mathbf{v}}$ is a Wasserstein matrix for $G^{\mathbf{v}}$.

Proof. Let $f : \mathbb{S} \to \mathbb{S}$ be a bounded and measurable quasilocal function, and let $x, z \in \mathbb{S}$ be configurations that differ at a single site $\operatorname{card}\{i \in I : x_i \neq z_i\} = 1$. Note that

$$\gamma^J_x f^J_x = (\gamma^J_x \otimes \delta_{x^{I\setminus J}})f, \qquad \gamma^J_z f^J_z = (\gamma^J_z \otimes \delta_{z^{I\setminus J}})f.$$

As $Q^J_{x,z}$ is a coupling of γ^J_x and γ^J_z by construction, the measure $Q^J_{x,z} \otimes \delta_{x^{I\setminus J}} \otimes \delta_{z^{I\setminus J}}$ is a coupling of $\gamma^J_x \otimes \delta_{x^{I\setminus J}}$ and $\gamma^J_z \otimes \delta_{z^{I\setminus J}}$. Thus Lemma C.3 yields

$$\begin{aligned}|\gamma^J_x f^J_x - \gamma^J_z f^J_z| &\leq \sum_{i\in I} \operatorname{osc}_i f\, (Q^J_{x,z} \otimes \delta_{x^{I\setminus J}} \otimes \delta_{z^{I\setminus J}})\eta_i \\ &= \sum_{i\in J} \operatorname{osc}_i f\, Q^J_{x,z}\eta_i + \sum_{i\in I\setminus J} \operatorname{osc}_i f\, \eta_i(x_i,z_i).\end{aligned}$$

In particular, we obtain

$$\begin{aligned}|G^{\mathbf{v}}f(x) - G^{\mathbf{v}}f(z)| &\leq \left(1 - \sum_{J\in\mathcal{J}} v_J\right)|f(x)-f(z)| + \sum_{J\in\mathcal{J}} v_J\,|\gamma^J_x f^J_x - \gamma^J_z f^J_z| \\ &\leq \left(1-\sum_{J\in\mathcal{J}} v_J\right)\sum_{i\in I}\operatorname{osc}_i f\,\eta_i(x_i,z_i) + \sum_{J\in\mathcal{J}} v_J\left(\sum_{i\in J}\operatorname{osc}_i f\, Q^J_{x,z}\eta_i + \sum_{i\in I\setminus J}\operatorname{osc}_i f\,\eta_i(x_i,z_i)\right) \\ &= \sum_{i\in I}\operatorname{osc}_i f\,\{1 - W^{\mathbf{v}}_{ii}\}\,\eta_i(x_i,z_i) + \sum_{i\in I}\operatorname{osc}_i f \sum_{J\in\mathcal{J}: i\in J} v_J\, Q^J_{x,z}\eta_i.\end{aligned}$$

Now suppose that $x^{I\setminus\{j\}} = z^{I\setminus\{j\}}$ (and $x \neq z$). Then by definition

$$\sum_{J\in\mathcal{J}: i\in J} v_J\, Q^J_{x,z}\eta_i \leq R^{\mathbf{v}}_{ij}\,\eta_j(x_j,v_j),$$

and we obtain

$$\frac{|G^{\mathbf{v}}f(x) - G^{\mathbf{v}}f(z)|}{\eta_j(x_j, z_j)} \leq \operatorname{osc}_j f \, \{1 - W^{\mathbf{v}}_{jj}\} + \sum_{i \in I} \operatorname{osc}_i f \, R^{\mathbf{v}}_{ij}.$$

Thus $V^{\mathbf{v}} = I - W^{\mathbf{v}} + R^{\mathbf{v}}$ satisfies Definition C.1. □

Using Lemmas C.4 and C.5, we can now apply Proposition C.2.

Corollary C.6. *Assume that γ^J is quasilocal for every $J \in \mathcal{J}$. Then*

$$|\rho f - \tilde{\rho} f| \leq \sum_{i,j \in I} \operatorname{osc}_i f \, N^{\mathbf{v}(n)}_{ij} \, a^{\mathbf{v}}_j + \sum_{i,j \in I} \operatorname{osc}_i f \, (I - W^{\mathbf{v}} + R^{\mathbf{v}})^n_{ij} \, (\rho \otimes \tilde{\rho}) \eta_j$$

for every $n \geq 1$ and bounded and measurable quasilocal function f, where

$$N^{\mathbf{v}(n)} := \sum_{k=0}^{n-1} (I - W^{\mathbf{v}} + R^{\mathbf{v}})^k$$

and the coefficients $(a^{\mathbf{v}}_j)_{j \in I}$ are defined by $a^{\mathbf{v}}_j := \sum_{J \in \mathcal{J} : j \in J} v_J \int^* \tilde{\rho}(dx) \, \hat{Q}^J_x \eta_j$.

Proof. Let $G = G^{\mathbf{v}}$, $\tilde{G} = \tilde{G}^{\mathbf{v}}$, $V = I - W^{\mathbf{v}} + R^{\mathbf{v}}$ in Proposition C.2. The requisite assumptions are verified by Lemmas C.4 and C.5, and it remains to show that there exists a coupling Q_x of G_x and $\tilde{G}_x$ such that $\int^* \tilde{\rho}(dx) \, Q_x \eta_j \leq a_j$ for every $j \in I$. But choosing

$$Q_x g := \left(1 - \sum_{J \in \mathcal{J}} v_J\right) g(x, x) + \sum_{J \in \mathcal{J}} v_J \int \hat{Q}^J_x(dz^J, d\tilde{z}^J) \, g(z^J x^{I \setminus J}, \tilde{z}^J x^{I \setminus J}),$$

it is easily verified that Q_x satisfies the necessary properties. □

In order for the construction of the Gibbs sampler to make sense, the weights v_J must be probabilities. This imposes the requirement $\sum_J v_J \leq 1$. If we were to assume that $\sum_J w_J \leq 1$, we could apply Corollary C.6 with $v_J = w_J$. Then assumption (6.1) guarantees that the second term in Corollary C.6 vanishes as $n \to \infty$, which yields

$$|\rho f - \tilde{\rho} f| \leq \sum_{i,j \in I} \operatorname{osc}_i f \, N_{ij} \, a_j \quad \text{with} \quad N := \sum_{k=0}^{\infty} (I - W + R)^k.$$

The proof of Theorem 6.4 would now be complete after we establish the identity

$$N = \sum_{k=0}^{\infty} (I - W + R)^k = \sum_{k=0}^{\infty} (W^{-1} R)^k \, W^{-1} = D W^{-1}.$$

This straightforward matrix identity will be proved in the next section. The assumption that the weights w_J are summable is restrictive, however, when I is infinite:

in Theorem 6.4 we only assume that $W_{ii} \le 1$ for all i, so we evidently cannot set $v_J = w_J$.

When the weights w_j are not summable, it is not natural to interpret them as probabilities. In this setting, a much more natural construction would be to consider a *continuous time* counterpart of the Gibbs sampler called *Glauber dynamics*. To define this process, one attaches to each region $J \in \mathcal{J}$ an independent Poisson process with rate w_J, and applies the transition kernel γ^J at every jump time of the corresponding Poisson process. Thus w_J does not represent the probability of selecting the region J in one time step, but rather the frequency with which region J is selected in continuous time. Once this process has been defined, one would choose the transition kernel G to be the transition semigroup of the continuous time process on any fixed time interval. Proceeding with this construction we expect, at least formally, to obtain Theorem 6.4 under the stated assumptions.

Unfortunately, there are nontrivial technical issues involved in implementing this approach: it is not evident *a priori* that the continuous time construction defines a well-behaved Markov semigroup, so that it is unclear when the above program can be made rigorous. The existence of a semigroup has typically been established under more restrictive assumptions than we have imposed in the present setting [36]. In order to circumvent such issues, we will proceed by an alternate route. Formally, the Glauber dynamics can be obtained by an appropriate scaling limit of discrete time Gibbs samplers. We will also utilize this scaling, but instead of applying Proposition C.2 to the limiting dynamics we will take the scaling limit directly in Corollary C.6. Thus, while our intuition comes from the continuous time setting, we avoid some technicalities inherent in the construction of the limit dynamics. Instead, we now face the problem of taking limits of powers of infinite matrices. The requisite matrix algebra will be worked out in the following section.

Remark C.7. *Let us briefly sketch how the previous results can be sharpened to obtain a* nonlinear *comparison theorem that could lead to sharper bounds in some situations. Assume for simplicity that* $\sum_J w_J \le 1$. *Then* $V = I - W + R$ *is a Wasserstein matrix for* G *by Lemma C.5. Writing out the definitions, this means* $\delta(Gf) \le \delta(f)V$ *where*

$$(\beta V)_j = \sum_{i \in I} \beta_i \sup_{\substack{x,z \in \mathbb{S}: \\ x^{I\setminus\{j\}} = z^{I\setminus\{j\}}}} \left\{ \mathbf{1}_{i=j} \left(1 - \sum_{J: i \in J} w_J \right) + \frac{1}{\eta_j(x_j, z_j)} \sum_{J: i \in J} w_J \, Q^J_{x,z} \eta_i \right\}$$

(here we interpret $\beta = (\beta_i)_{i \in I}$ *and* $\delta(f) = (\mathrm{osc}_i f)_{i \in I}$ *as row vectors). However, from the proof of Lemma C.5 we even obtain the sharper bound* $\delta(Gf) \le \mathsf{V}[\delta(f)]$ *where*

$$\mathsf{V}[\beta]_j := \sup_{\substack{x,z \in \mathbb{S}: \\ x^{I\setminus\{j\}} = z^{I\setminus\{j\}}}} \sum_{i \in I} \beta_i \left\{ \mathbf{1}_{i=j} \left(1 - \sum_{J: i \in J} w_J \right) + \frac{1}{\eta_j(x_j, z_j)} \sum_{J: i \in J} w_J \, Q^J_{x,z} \eta_i \right\}$$

is defined with the supremum over configurations outside the sum. The nonlinear operator V *can now be used much in the same way as the Wasserstein matrix* V. *In*

particular, following the identical proof as for Proposition C.2, we immediately obtain

$$|\rho f - \tilde{\rho} f| \le \sum_{j \in I} \sum_{k=0}^{n-1} \mathsf{V}^k[\delta(f)]_j \int^* \tilde{\rho}(dx)\, Q_x \eta_j + \sum_{j \in I} \mathsf{V}^n[\delta(f)]_j\, (\rho \otimes \tilde{\rho}) \eta_j,$$

where V^k denotes the kth iterate of the nonlinear operator V. Proceeding along these lines, one can develop nonlinear comparison theorems under Dobrushin-Shlosman type conditions (see the discussion in Section 6.3.2). The nonlinear expressions are somewhat difficult to handle, however, and we do not develop this idea further in this thesis.

C.3 Proof of Theorem 6.4

Throughout this section, we work under the assumptions of Theorem 6.4. The main idea of the proof is the following continuous scaling limit of Corollary C.6.

Proposition C.8. *Let $t > 0$. Define the matrices*

$$N := \sum_{k=0}^{\infty} (I - W + R)^k, \qquad V^{[t]} := \sum_{k=0}^{\infty} \frac{t^k e^{-t}}{k!} (I - W + R)^k.$$

Then we have, under the assumptions of Theorem 6.4,

$$|\rho f - \tilde{\rho} f| \le \sum_{i,j \in I} \mathrm{osc}_i f\, N_{ij}\, a_j + \sum_{i,j \in I} \mathrm{osc}_i f\, V^{[t]}_{ij}\, (\rho \otimes \tilde{\rho}) \eta_j$$

for every bounded and measurable quasilocal function f such that $\mathrm{osc}_i f < \infty$ for all $i \in I$.

Proof. Without loss of generality, we will assume throughout the proof that f is a local function (so that only finitely many $\mathrm{osc}_i f$ are nonzero). The extension to quasilocal f follows readily by applying the local result to f^J_x and letting $J \uparrow I$ as in the proof of Lemma C.3.

As the cover $\mathcal{J}$ is at most countable (because $\mathcal{I}$ is countable), we can enumerate its elements arbitrarily as $\mathcal{J} = \{J_1, J_2, \ldots\}$. Define the weights $\mathbf{v}^r = (v^r_J)_{J \in \mathcal{J}}$ as

$$v^r_J := \begin{cases} w_J & \text{when } J = J_k \text{ for } k \le r, \\ 0 & \text{otherwise.} \end{cases}$$

For every $r \in \mathbb{N}$, the weight vector $u\mathbf{v}^r$ evidently satisfies $\sum_J u v^r_J \le 1$ for all $u > 0$ sufficiently small (depending on r). The main idea of the proof is to apply Corollary C.6 to the weight vector $\mathbf{v} = (t/n)\mathbf{v}^r$, then let $n \to \infty$, and finally $r \to \infty$.

Let us begin by considering the second term in Corollary C.6. We can write

$$\begin{aligned} (I - W^{(t/n)\mathbf{v}^r} + R^{(t/n)\mathbf{v}^r})^n &= \left(\left(1 - \frac{t}{n}\right) I + \frac{t}{n} (I - W^{\mathbf{v}^r} + R^{\mathbf{v}^r}) \right)^n \\ &= \sum_{k=0}^{n} \binom{n}{k} \left(1 - \frac{t}{n}\right)^{n-k} \left(\frac{t}{n}\right)^k (I - W^{\mathbf{v}^r} + R^{\mathbf{v}^r})^k \\ &= \mathbf{E}\, (I - W^{\mathbf{v}^r} + R^{\mathbf{v}^r})^{Z_n}, \end{aligned}$$

where we defined the Binomial random variables $Z_n \sim \mathrm{Bin}(n, t/n)$. The random variables Z_n converge weakly as $n \to \infty$ to the Poisson random variable $Z_\infty \sim \mathrm{Pois}(t)$. To take the limit of the above expectation, we need a simple estimate that will be useful in the sequel.

Lemma C.9. *Let $(c_j)_{j\in I}$ be any nonnegative vector. Then*

$$\sum_{j\in I}(I - W^{\mathbf{v}^r} + R^{\mathbf{v}^r})^k_{ij}\, c_j \le 2^k \max_{0\le\ell\le k}\sum_{j\in I}(I - W + R)^\ell_{ij}\, c_j$$

for every $i \in I$ and $k \ge 0$.

Proof. As $R^{\mathbf{v}}$ is nondecreasing in $\mathbf{v}$ we obtain the elementwise estimate

$$I - W^{\mathbf{v}^r} + R^{\mathbf{v}^r} \le I + R \le I + (I - W + R),$$

where we have used $W_{ii} \le 1$. We therefore have

$$\sum_{j\in I}(I - W^{\mathbf{v}^r} + R^{\mathbf{v}^r})^k_{ij}\, c_j \le \sum_{j\in I}(I + \{I - W + R\})^k_{ij}\, c_j = \sum_{\ell=0}^{k}\binom{k}{\ell}\sum_{j\in I}(I - W + R)^\ell_{ij}\, c_j,$$

and the proof is easily completed. □

Define the random variables

$$X_n = g(Z_n) \quad \text{with} \quad g(k) = \sum_{i,j\in I}\mathrm{osc}_i f\,(I - W^{\mathbf{v}^r} + R^{\mathbf{v}^r})^k_{ij}\,(\rho\otimes\tilde\rho)\eta_j.$$

Then $X_n \to X_\infty$ weakly by the continuous mapping theorem. On the other hand, applying Lemma C.9 with $c_j = (\rho\otimes\tilde\rho)\eta_j$ we estimate $g(k) \le C2^k$ for some finite constant $C < \infty$ and all $k \ge 0$, where we have used assumption (6.1) and that f is local. As

$$\limsup_{u\to\infty}\sup_{n\ge1}\mathbf{E}(2^{Z_n}\mathbf{1}_{2^{Z_n}\ge u}) \le \lim_{u\to\infty} u^{-1}\sup_{n\ge1}\mathbf{E}\,4^{Z_n} = \lim_{u\to\infty} u^{-1}e^{3t} = 0,$$

it follows that the random variables $(X_n)_{n\ge1}$ are uniformly integrable. We therefore conclude that $\mathbf{E}\,X_n \to \mathbf{E}\,X_\infty$ as $n \to \infty$ (cf. [31, Lemma 4.11]). In particular,

$$\lim_{n\to\infty}\sum_{i,j\in I}\mathrm{osc}_i f\,(I - W^{(t/n)\mathbf{v}^r} + R^{(t/n)\mathbf{v}^r})^n_{ij}\,(\rho\otimes\tilde\rho)\eta_j = \sum_{i,j\in I}\mathrm{osc}_i f\, V^{r[t]}_{ij}\,(\rho\otimes\tilde\rho)\eta_j,$$

where

$$V^{r[t]} = \sum_{k=0}^{\infty}\frac{t^k e^{-t}}{k!}(I - W^{\mathbf{v}^r} + R^{\mathbf{v}^r})^k.$$

We now let $r \to \infty$. Note that $W^{\mathbf{v}^r} \uparrow W$ and $R^{\mathbf{v}^r} \uparrow R$ elementwise and, arguing as in the proof of Lemma C.9, we have $I - W^{\mathbf{v}^r} + R^{\mathbf{v}^r} \le I + (I - W + R)$ elementwise where

$$\sum_{k=0}^{\infty}\sum_{i,j\in I}\frac{t^k e^{-t}}{k!}\,\mathrm{osc}_i f\,\{I+(I-W+R)\}^k_{ij}\,(\rho\otimes\tilde\rho)\eta_j \le e^t\sup_{\ell\ge0}\sum_{i,j\in I}\mathrm{osc}_i f\,(I-W+R)^\ell_{ij}\,(\rho\otimes\tilde\rho)\eta_j$$

is finite by assumption (6.1) and as f is local. We therefore obtain

$$\lim_{r\to\infty}\lim_{n\to\infty}\sum_{i,j\in I}\mathrm{osc}_i f\,(I - W^{(t/n)\mathbf{v}^r} + R^{(t/n)\mathbf{v}^r})^n_{ij}\,(\rho\otimes\tilde\rho)\eta_j = \sum_{i,j\in I}\mathrm{osc}_i f\,V^{[t]}_{ij}\,(\rho\otimes\tilde\rho)\eta_j$$

by dominated convergence. That is, the second term in Corollary C.6 with $\mathbf{v} = (t/n)\mathbf{v}^r$ converges as $n\to\infty$ and $r\to\infty$ to the second term in statement of the present result.

It remains to establish the corresponding conclusion for the first term in Corollary C.6, which proceeds much along the same lines. We begin by noting that

$$\begin{aligned}\frac{1}{n}\sum_{k=0}^{n-1}(I - W^{(t/n)\mathbf{v}^r} + R^{(t/n)\mathbf{v}^r})^k &= \frac{1}{n}\sum_{k=0}^{n-1}\left(\left(1-\frac{t}{n}\right)I + \frac{t}{n}(I - W^{\mathbf{v}^r} + R^{\mathbf{v}^r})\right)^k\\ &= \frac{1}{n}\sum_{k=0}^{n-1}\sum_{\ell=0}^{k}\binom{k}{\ell}\left(1-\frac{t}{n}\right)^{k-\ell}\left(\frac{t}{n}\right)^{\ell}(I - W^{\mathbf{v}^r} + R^{\mathbf{v}^r})^\ell\\ &= \sum_{\ell=0}^{n-1}p^{(n)}_\ell(I - W^{\mathbf{v}^r} + R^{\mathbf{v}^r})^\ell,\end{aligned}$$

where we have defined

$$p^{(n)}_\ell = \frac{1}{n}\sum_{k=\ell}^{n-1}\binom{k}{\ell}\left(1-\frac{t}{n}\right)^{k-\ell}\left(\frac{t}{n}\right)^{\ell} = \frac{1}{t}\int_{\ell t/n}^{t}\binom{\lfloor sn/t\rfloor}{\ell}\left(1-\frac{t}{n}\right)^{\lfloor sn/t\rfloor-\ell}\left(\frac{t}{n}\right)^{\ell}ds$$

for $\ell < n$. An elementary computation yields

$$\sum_{\ell=0}^{n-1}p^{(n)}_\ell = 1 \qquad\text{and}\qquad p^{(n)}_\ell \xrightarrow{n\to\infty} p^{(\infty)}_\ell = \frac{1}{t}\int_0^t\frac{s^\ell e^{-s}}{\ell!}\,ds.$$

We can therefore introduce $\{0,1,\ldots\}$-valued random variables Y_n with $\mathbf{P}(Y_n=\ell) = p^{(n)}_\ell$ for $\ell < n$, and we have shown above that $Y_n\to Y_\infty$ weakly and that

$$\frac{1}{n}\sum_{k=0}^{n-1}(I - W^{(t/n)\mathbf{v}^r} + R^{(t/n)\mathbf{v}^r})^k = \mathbf{E}\,(I - W^{\mathbf{v}^r} + R^{\mathbf{v}^r})^{Y_n}.$$

The first term in Corollary C.6 with $\mathbf{v} = (t/n)\mathbf{v}^r$ can be written as

$$\sum_{i,j\in I}\mathrm{osc}_i f\sum_{k=0}^{n-1}(I - W^{(t/n)\mathbf{v}^r} + R^{(t/n)\mathbf{v}^r})^k_{ij}\,a^{(t/n)\mathbf{v}^r}_j = t\,\mathbf{E}\,h(Y_n),$$

where we have defined

$$h(k) = \sum_{i,j\in I}\mathrm{osc}_i f\,(I - W^{\mathbf{v}^r} + R^{\mathbf{v}^r})^k_{ij}\,a^{\mathbf{v}^r}_j.$$

We now proceed essentially as above. We can assume without loss of generality that

$$\sup_{\ell\geq 0}\sum_{i,j\in I}\operatorname{osc}_i f\,(I-W+R)^{\ell}_{ij}\,a_j<\infty,$$

as otherwise the right-hand side in the statement of the present result is infinite and the estimate is trivial. It consequently follows from Lemma C.9 that $h(k)\leq C2^k$ for some finite constant $C<\infty$ and all $k\geq 0$. A similar computation as was done above shows that $(h(Y_n))_{n\geq 0}$ is uniformly integrable, and therefore $\mathbf{E}\,h(Y_n)\to\mathbf{E}\,h(Y_\infty)$. In particular, the first term in Corollary C.6 with $\mathbf{v}=(t/n)\mathbf{v}^r$ converges as $n\to\infty$ to

$$\lim_{n\to\infty}\sum_{i,j\in I}\operatorname{osc}_i f\sum_{k=0}^{n-1}(I-W^{(t/n)\mathbf{v}^r}+R^{(t/n)\mathbf{v}^r})^k_{ij}\,a_j^{(t/n)\mathbf{v}^r}=\sum_{i,j\in I}\operatorname{osc}_i f\,N^r_{ij}\,a_j^{\mathbf{v}^r},$$

where

$$N^r=\sum_{k=0}^{\infty}\int_0^t\frac{s^ke^{-s}}{k!}\,ds\;(I-W^{\mathbf{v}^r}+R^{\mathbf{v}^r})^k.$$

Similarly, letting $r\to\infty$ and repeating exactly the arguments used above for the second term of Corollary C.6, we obtain by dominated convergence

$$\lim_{r\to\infty}\lim_{n\to\infty}\sum_{i,j\in I}\operatorname{osc}_i f\sum_{k=0}^{n-1}(I-W^{(t/n)\mathbf{v}^r}+R^{(t/n)\mathbf{v}^r})^k_{ij}\,a_j^{(t/n)\mathbf{v}^r}=\sum_{i,j\in I}\operatorname{osc}_i f\,\tilde{N}_{ij}\,a_j,$$

where

$$\tilde{N}=\sum_{k=0}^{\infty}\int_0^t\frac{s^ke^{-s}}{k!}\,ds\;(I-W+R)^k.$$

To conclude, we have shown that applying Corollary C.6 to the weight vector $\mathbf{v}=(t/n)\mathbf{v}^r$ and taking the limit as $n\to\infty$ and $r\to\infty$, respectively, yields the estimate

$$|\rho f-\tilde{\rho}f|\leq\sum_{i,j\in I}\operatorname{osc}_i f\,\tilde{N}_{ij}\,a_j+\sum_{i,j\in I}\operatorname{osc}_i f\,V^{[t]}_{ij}\,(\rho\otimes\tilde{\rho})\eta_j.$$

It remains to note that $t^ke^{-t}/k!$ is the density of a Gamma distribution (with shape $k+1$ and scale 1), so $\int_0^t s^ke^{-s}/k!\,ds\leq 1$ and thus $\tilde{N}\leq N$ elementwise. □

We can now complete the proof of Theorem 6.4.

Proof of Theorem 6.4. Once again, we will assume without loss of generality that f is a local function (so that only finitely many $\operatorname{osc}_i f$ are nonzero). The extension to quasilocal f follows readily by localization as in the proof of Lemma C.3.

We begin by showing that the second term in Proposition C.8 vanishes as $t \to \infty$. Indeed, for any $n \geq 0$, we can evidently estimate the second term as

$$\begin{aligned}
&\sum_{k=0}^{\infty} \frac{t^k e^{-t}}{k!} \sum_{i,j \in I} \operatorname{osc}_i f \, (I - W + R)^k_{ij} \, (\rho \otimes \tilde{\rho}) \eta_j \\
&\qquad \leq \sup_{\ell \geq 0} \sum_{i,j \in I} \operatorname{osc}_i f \, (I - W + R)^\ell_{ij} \, (\rho \otimes \tilde{\rho}) \eta_j \sum_{k=0}^{n} \frac{t^k e^{-t}}{k!} \\
&\qquad\quad + \sup_{\ell > n} \sum_{i,j \in I} \operatorname{osc}_i f \, (I - W + R)^\ell_{ij} \, (\rho \otimes \tilde{\rho}) \eta_j .
\end{aligned}$$

By assumption (6.1) and as f is local, the two terms on the right vanish as $t \to \infty$ and $n \to \infty$, respectively. Thus second term in Proposition C.8 vanishes as $t \to \infty$.

We have now proved the estimate

$$|\rho f - \tilde{\rho} f| \leq \sum_{i,j \in I} \operatorname{osc}_i f \, N_{ij} \, a_j .$$

To complete the proof of Theorem 6.4, it remains to establish the identity $N = DW^{-1}$. This is an exercise in matrix algebra. By the definition of the matrix product, we have

$$(I - W + R)^p = \sum_{k=0}^{p} \sum_{\substack{n_0, \ldots, n_k \geq 0 \\ n_0 + \cdots + n_k = p - k}} (I - W)^{n_k} R \cdots (I - W)^{n_1} R (I - W)^{n_0} .$$

We can therefore write

$$\begin{aligned}
&\sum_{p=0}^{\infty} (I - W + R)^p \\
&\quad = \sum_{k=0}^{\infty} \sum_{n_0, \ldots, n_k \geq 0} \sum_{p=0}^{\infty} \mathbf{1}_{n_0 + \cdots + n_k = p - k} \mathbf{1}_{k \leq p} (I - W)^{n_k} R \cdots (I - W)^{n_1} R (I - W)^{n_0} \\
&\quad = \sum_{k=0}^{\infty} \sum_{n_0, \ldots, n_k \geq 0} (I - W)^{n_k} R \cdots (I - W)^{n_1} R (I - W)^{n_0} \\
&\quad = \sum_{k=0}^{\infty} (W^{-1} R)^k W^{-1} ,
\end{aligned}$$

where we have used that $W^{-1} = \sum_{n=0}^{\infty} (I - W)^n$ as W is diagonal with $0 < W_{ii} \leq 1$. □

C.4 Proof of Corollary 6.8

Note that $\sup_i W_{ii} < \infty$ in all parts of Corollary 6.8 (either by assumption or as $\operatorname{card} I < \infty$). Moreover, it is easily seen that all parts of Corollary 6.8 as well as the

conclusion of Theorem 6.4 are unchanged if all the weights are multiplied by the same constant. We may therefore assume without loss of generality that $\sup_i W_{ii} \le 1$.

Next, we note that as ρ and $\tilde\rho$ are tempered, we have

$$\sup_{i\in I} (\rho\otimes\tilde\rho)\eta_i \le \sup_{i\in I} \rho\,\eta_i(\,\cdot\,, x_i^\star) + \sup_{i\in I} \tilde\rho\,\eta_i(x_i^\star,\,\cdot\,) < \infty$$

by the triangle inequality. To verify (6.1), it therefore suffices to show that

$$\lim_{k\to\infty} \sum_{j\in I} (I - W + R)_{ij}^k = 0 \quad \text{for all } i \in I. \tag{C.1}$$

We now proceed to verify this condition in the different cases of Corollary 6.8.

Proof of Corollary 6.8(1). It was shown at the end of the proof of Theorem 6.4 that

$$\sum_{k=0}^\infty (I - W + R)^k = \sum_{k=0}^\infty (W^{-1}R)^k W^{-1} = DW^{-1}.$$

As W^{-1} has finite entries, $D < \infty$ certainly implies that $(I-W+R)^k \to 0$ as $k\to\infty$ elementwise. But this trivially yields (C.1) when $\operatorname{card} I < \infty$. □

Proof of Corollary 6.8(2). Note that we can write

$$D = \sum_{k=0}^\infty (W^{-1}R)^k = \sum_{p=0}^{n-1} (W^{-1}R)^p \sum_{k=0}^\infty (W^{-1}R)^{nk}.$$

Therefore, if $R < \infty$ and $\|(W^{-1}R)^n\| < 1$, we can estimate

$$\|D\| \le \left\| \sum_{p=0}^{n-1} (W^{-1}R)^p \right\| \sum_{k=0}^\infty \|(W^{-1}R)^n\|^k < \infty.$$

Thus $D < \infty$ and we conclude by the previous part. □

Proof of Corollary 6.8(3). We give a simple probabilistic proof (a more complicated matrix-analytic proof could be given along the lines of [14, Theorem 3.21]). Let $P = W^{-1}R$. As $\|P\|_\infty < 1$, the infinite matrix P is substochastic. Thus P is the transition probability matrix of a killed Markov chain $(X_n)_{n\ge0}$ such that $\mathbf{P}(X_n = j | X_{n-1} = i) = P_{ij}$ and $\mathbf{P}(X_n \text{ is dead} | X_{n-1} = i) = 1 - \sum_j P_{ij}$ (once the chain dies, it stays dead). Denote by $\zeta = \inf\{n : X_n \text{ is dead}\}$ the killing time of the chain. Then we obtain

$$\mathbf{P}(\zeta > n | X_0 = i) = \mathbf{P}(X_n \text{ is not dead} | X_0 = i) = \sum_{j\in I} P_{ij}^n \le \|P^n\|_\infty \le \|P\|_\infty^n.$$

Therefore, as $\|P\|_\infty < 1$, we find by letting $n\to\infty$ that $\mathbf{P}(\zeta = \infty | X_0 = i) = 0$. That is, the chain dies eventually with unit probability for any initial condition.

Now define $\tilde{P} = I - W + R = I - W + WP$. As $\sup_i W_{ii} \le 1$, the matrix $\tilde{P}$ is also substochastic and corresponds to the following transition mechanism. If $X_{n-1} = i$, then at time n we flip a biased coin that comes up heads with probability W_{ii}. In case of heads we make a transition according to the matrix P, but in case of tails we leave the current state unchanged. From this description, it is evident that we can construct a Markov chain $(\tilde{X}_n)_{n\ge 0}$ with transition matrix $\tilde{P}$ by modifying the chain $(X_n)_{n\ge 0}$ as follows. Conditionally on $(X_n)_{n\ge 0}$, draw independent random variables $(\xi_n)_{n\ge 0}$ such that ξ_n is geometrically distributed with parameter $W_{X_n X_n}$. Now define the process $(\tilde{X}_n)_{n\ge 0}$ such that it stays in state X_0 for the first ξ_0 time steps, then is in state X_1 for the next ξ_1 time steps, etc. By construction, the resulting process is Markov with transition matrix $\tilde{P}$. Moreover, as $\zeta < \infty$ a.s., we have $\tilde{\zeta} := \inf\{n : \tilde{X}_n \text{ is dead}\} < \infty$ a.s. also. Thus

$$\lim_{n\to\infty} \sum_{j\in I} (I - W + R)^n_{ij} = \lim_{n\to\infty} \mathbf{P}(\tilde{\zeta} > n | X_0 = i) = 0$$

for every $i \in I$. We have therefore established (C.1). □

Proof of Corollary 6.8(4). We begin by writing as above

$$\sum_{k=0}^{\infty} (I - W + R)^k = \sum_{k=0}^{\infty} (W^{-1}R)^k W^{-1} = \sum_{k=0}^{\infty} W^{-1}(RW^{-1})^k,$$

where the last identity is straightforward. Arguing as in Corollary 6.8(2), we obtain

$$\begin{aligned} W_{ii} \sum_{k=0}^{\infty} \sum_{j\in I} (I - W + R)^k_{ij} = \sum_{j\in I} \sum_{k=0}^{\infty} (RW^{-1})^k_{ij} &\le \left\| \sum_{k=0}^{\infty} (RW^{-1})^k \right\|_\infty \\ &\le \sum_{p=0}^{n-1} \|RW^{-1}\|^p_\infty \sum_{k=0}^{\infty} \|(RW^{-1})^n\|^k_\infty < \infty. \end{aligned}$$

It follows immediately that (C.1) holds. □

Proof of Corollary 6.8(5). Note that

$$\sum_{j\in I} (RW^{-1})^k_{ij} \|\eta_j\| \le \|(RW^{-1})^k\|_1 \sum_{j\in I} \|\eta_j\| \le \|RW^{-1}\|^k_1 \sum_{j\in I} \|\eta_j\|.$$

Thus $\sum_j \|\eta_j\| < \infty$ and $\|RW^{-1}\|_1 < 1$ yield

$$\sum_{k=0}^{\infty} \sum_{j\in I} (I - W + R)^k_{ij} \|\eta_j\| = W^{-1}_{ii} \sum_{k=0}^{\infty} \sum_{j\in I} (RW^{-1})^k_{ij} \|\eta_j\| < \infty,$$

which evidently implies

$$\lim_{k\to\infty} \sum_{j\in I} (I - W + R)^k_{ij} (\rho \otimes \tilde{\rho}) \eta_j = 0 \quad \text{for all } i \in I.$$

We have therefore established (6.1). □

Proof of Corollary 6.8(6). Let $r = \sup\{m(i,j) : R_{ij} > 0\}$ (which is finite by assumption), and choose $\beta > 0$ such that $\|RW^{-1}\|_1 < e^{-\beta r}$. Then we can estimate

$$\|RW^{-1}\|_{1,\beta m} := \sup_{j\in I} \sum_{i\in I} e^{\beta m(i,j)} (RW^{-1})_{ij} \le e^{\beta r} \|RW^{-1}\|_1 < 1.$$

As m is a pseudometric, it satisfies the triangle inequality and it is therefore easily seen that $\|\cdot\|_{1,\beta m}$ is a matrix norm. In particular, we can estimate

$$e^{\beta m(i,j)} (RW^{-1})^n_{ij} \le \|(RW^{-1})^n\|_{1,\beta m} \le \|RW^{-1}\|^n_{1,\beta m}$$

for every $i, j \in I$. But then

$$\|(RW^{-1})^n\|_\infty = \sup_{i\in I} \sum_{j\in I} (RW^{-1})^n_{ij} \le \|RW^{-1}\|^n_{1,\beta m} \sup_{i\in I} \sum_{j\in I} e^{-\beta m(i,j)} < \infty$$

for all n. We therefore have $\|RW^{-1}\|_\infty < \infty$, and we can choose n sufficiently large that $\|(RW^{-1})^n\|_\infty < 1$. The conclusion now follows from Corollary 6.8(4). □

C.5 Proof of Theorem 6.12

In the case of one-sided local updates, the measure $\rho_{\le k}$ is γ^J-invariant for $\tau(J) = k$ (but not for $\tau(J) < k$). The proof of Theorem 6.12 therefore proceeds by induction on k. In each stage of the induction, we apply the logic of Theorem 6.4 to the partial local updates $(\gamma^J)_{J\in\mathfrak{J}:\tau(J)=k}$, and use the induction hypothesis to estimate the remainder term.

Throughout this section, we work in the setting of Theorem 6.12. Define

$$I_{\le k} := \{i \in I : \tau(i) \le k\}, \qquad I_k := \{i \in I : \tau(i) = k\}.$$

Note that we can assume without loss of generality that $R_{ij} = 0$ whenever $\tau(j) > \tau(i)$. Indeed, the local update rule γ^J_x does not depend on x_j for $\tau(j) > \tau(J)$, so we can trivially choose the coupling $Q^J_{x,z}$ for $x^{I\setminus\{j\}} = z^{I\setminus\{j\}}$ such that $Q^J_{x,z}\eta_i = 0$ for all $i \in J$. On the other hand, the choice $R_{ij} = 0$ evidently yields the smallest bound in Theorem 6.12. In the sequel, we will always assume that $R_{ij} = 0$ whenever $\tau(j) > \tau(i)$.

The key induction step is formalized by the following result.

Proposition C.10. *Assume (6.1). Let $(\beta_i)_{i\in I_{\le k-1}}$ be nonnegative weights such that*

$$|\rho_{\le k-1} g - \tilde\rho_{\le k-1} g| \le \sum_{i\in I_{\le k-1}} \mathrm{osc}_i g\, \beta_i$$

for every bounded measurable quasilocal function g on $\mathbb{S}_{\le k-1}$ so that $\mathrm{osc}_i g < \infty$ $\forall i$. Then

$$|\rho_{\le k} f - \tilde\rho_{\le k} f| \le \sum_{j\in I_{\le k-1}} \left\{ \mathrm{osc}_j f + \sum_{i,l\in I_k} \mathrm{osc}_i f\, D_{il}\, (W^{-1}R)_{lj} \right\} \beta_j + \sum_{i,j\in I_k} \mathrm{osc}_i f\, D_{ij}\, W^{-1}_{jj} a_j$$

for every bounded measurable quasilocal function f on $\mathbb{S}_{\le k}$ so that $\mathrm{osc}_i f < \infty$ $\forall i$.

Proof. We fix throughout the proof a bounded and measurable local function $f : \mathbb{S}_{\leq k} \to \mathbb{R}$ such that $\mathrm{osc}_i f < \infty$ for all $i \in I_{\leq k}$. The extension of the conclusion to quasilocal functions f follows readily by localization as in the proof of Lemma C.3.

We denote by $G^{\mathbf{v}}$ and $\tilde{G}^{\mathbf{v}}$ the Gibbs samplers as defined in Section C.2. Let us enumerate the partial cover $\{J \in \mathcal{J} : \tau(J) = k\}$ as $\{J_1, J_2, \ldots\}$, and define the weights $\mathbf{v}^r$ as in the proof of Proposition C.8. By the definition of the one-sided local update rule, $\rho_{\leq k}$ is $G^{u\mathbf{v}^r}$-invariant and $\tilde{\rho}_{\leq k}$ is $\tilde{G}^{u\mathbf{v}^r}$-invariant for every r, u such that $\sum_J u v_J^r \leq 1$. Thus

$$|\rho_{\leq k} f - \tilde{\rho}_{\leq k} f| \leq \sum_{i,j \in I_{\leq k}} \mathrm{osc}_i f \, N_{ij}^{u\mathbf{v}^r(n)} \, a_j^{u\mathbf{v}^r} + |\rho_{\leq k}(G^{u\mathbf{v}^r})^n f - \tilde{\rho}_{\leq k}(G^{u\mathbf{v}^r})^n f|$$

as in the proof of Corollary C.6, with the only distinction that we refrain from using the Wasserstein matrix to expand the second term in the proof of Proposition C.2. We now use the induction hypothesis to obtain an improved estimate for the second term.

Lemma C.11. *We can estimate*

$$|\rho_{\leq k} g - \tilde{\rho}_{\leq k} g| \leq \sum_{i \in I_{\leq k-1}} \mathrm{osc}_i g \, \beta_i + 3 \sum_{i \in I_k} \mathrm{osc}_i g \, (\rho \otimes \tilde{\rho}) \eta_i$$

for any bounded and measurable quasilocal function $g : \mathbb{S}_{\leq k} \to \mathbb{R}$ such that $\mathrm{osc}_i g < \infty$ $\forall i$.

Proof. For any $x \in \mathbb{S}_{\leq k}$ we can estimate

$$|\rho_{\leq k} g - \tilde{\rho}_{\leq k} g| \leq |\rho_{\leq k-1} \hat{g}_x - \tilde{\rho}_{\leq k-1} \hat{g}_x| + |\rho_{\leq k}(g - \hat{g}_x)| + |\tilde{\rho}_{\leq k}(g - \hat{g}_x)|,$$

where we defined $\hat{g}_x(z) := g(z^{I_{\leq k-1}} x^{I_k})$. By Lemma C.3 we have

$$|g(z) - \hat{g}_x(z)| \leq \sum_{i \in I_k} \mathrm{osc}_i g \, \eta_i(z_i, x_i).$$

We can therefore estimate using the induction hypothesis and the triangle inequality

$$|\rho_{\leq k} g - \tilde{\rho}_{\leq k} g| \leq \sum_{i \in I_{\leq k-1}} \mathrm{osc}_i g \, \beta_i + \sum_{i \in I_k} \mathrm{osc}_i g \, \{\rho \eta_i(\,\cdot\,, \tilde{x}_i) + \eta_i(\tilde{x}_i, x_i) + \tilde{\rho} \eta_i(\,\cdot\,, x_i)\}$$

for all $x, \tilde{x} \in \mathbb{S}_{\leq k}$. Now integrate this expression with respect to $\rho(dx)\,\tilde{\rho}(d\tilde{x})$. □

To lighten the notation somewhat we will write $\mathbf{v} = u\mathbf{v}^r$ until further notice. Note that by construction $a_j^{\mathbf{v}} = 0$ whenever $\tau(j) < k$, while $R_{ij}^{\mathbf{v}} = 0$ whenever $\tau(j) > \tau(i)$ by assumption. Thus we obtain using Lemma C.11 and Lemma C.5

$$\begin{aligned} |\rho_{\leq k} f - \tilde{\rho}_{\leq k} f| \leq \sum_{i,j \in I_k} \mathrm{osc}_i f \, N_{ij}^{\mathbf{v}(n)} \, a_j^{\mathbf{v}} + 3 \sum_{i,j \in I_k} \mathrm{osc}_i f \, (I - W^{\mathbf{v}} + R^{\mathbf{v}})_{ij}^n \, (\rho \otimes \tilde{\rho}) \eta_j \\ + \sum_{i \in I_{\leq k}} \sum_{j \in I_{\leq k-1}} \mathrm{osc}_i f \, (I - W^{\mathbf{v}} + R^{\mathbf{v}})_{ij}^n \, \beta_j, \end{aligned}$$

provided that $\sum_i \operatorname{osc}_i f\,(I - W^{\mathbf{v}} + R^{\mathbf{v}})^n_{ij} < \infty$ for all j.

Next, note that as $v_J = 0$ for $\tau(J) < k$, we have $R^{\mathbf{v}}_{ij} = W^{\mathbf{v}}_{ij} = 0$ for $i \in I_{\leq k-1}$. Thus

$$V^{\mathbf{v}} = I - W^{\mathbf{v}} + R^{\mathbf{v}} = \begin{pmatrix} \check{V}^{\mathbf{v}} & \check{R}^{\mathbf{v}} \\ 0 & I \end{pmatrix},$$

where $\check{V}^{\mathbf{v}} := (V^{\mathbf{v}}_{ij})_{i,j\in I_k}$ and $\check{R}^{\mathbf{v}} := (R^{\mathbf{v}}_{ij})_{i\in I_k, j\in I_{\leq k-1}}$. In particular,

$$(I - W^{\mathbf{v}} + R^{\mathbf{v}})^n = \begin{pmatrix} (\check{V}^{\mathbf{v}})^n & \sum_{k=0}^{n-1} (\check{V}^{\mathbf{v}})^k \check{R}^{\mathbf{v}} \\ 0 & I \end{pmatrix}.$$

Moreover, as $R^{\mathbf{v}}_{ij} = 0$ whenever $\tau(j) > \tau(i)$, we evidently have $(\check{V}^{\mathbf{v}})^k_{ij} = (V^{\mathbf{v}})^k_{ij}$ for $i, j \in I_k$. Substituting into the above expression, we obtain

$$\begin{aligned} |\rho_{\leq k} f - \tilde{\rho}_{\leq k} f| \leq \sum_{i,j\in I_k} \operatorname{osc}_i f\, N^{\mathbf{v}(n)}_{ij}\, a^{\mathbf{v}}_j + 3 \sum_{i,j\in I_k} \operatorname{osc}_i f\, (I - W^{\mathbf{v}} + R^{\mathbf{v}})^n_{ij}\, (\rho \otimes \tilde{\rho})\eta_j \\ + \sum_{j\in I_{\leq k-1}} \left\{ \operatorname{osc}_j f + \sum_{i,l\in I_k} \operatorname{osc}_i f\, N^{\mathbf{v}(n)}_{il}\, R^{\mathbf{v}}_{lj} \right\} \beta_j \end{aligned}$$

provided that $\sum_i \operatorname{osc}_i f\,(I - W^{\mathbf{v}} + R^{\mathbf{v}})^n_{ij} < \infty$ for all j. But the latter is easily verified using (6.1) and Lemma C.9, as f is local and $\operatorname{osc}_i f < \infty$ for all i by assumption.

The remainder of the proof now proceeds precisely as in the proof of Proposition C.8 and Theorem 6.4. We set $\mathbf{v} = (t/n)\mathbf{v}^r$, let $n \to \infty$ and then $r \to \infty$. The arguments for the first two terms are identical to the proof of Proposition C.8, while the argument for the third term is essentially identical to the argument for the first term. The proof is then completed as in the proof of Theorem 6.4. We leave the details for the reader. □

We now proceed to complete the proof of Theorem 6.12.

Proof of Theorem 6.12. Consider first the case that $k_- := \inf_{i\in I} \tau(i) > -\infty$. In this setting, we say that the comparison theorem holds for a given $k \geq k_-$ if we have

$$|\rho_{\leq k} f - \tilde{\rho}_{\leq k} f| \leq \sum_{i,j\in I_{\leq k}} \operatorname{osc}_i f\, D_{ij}\, W^{-1}_{jj} a_j$$

for every bounded measurable quasilocal function f on $\mathbb{S}_{\leq k}$ such that $\operatorname{osc}_i f < \infty\ \forall i$. We can evidently apply Theorem 6.4 to show that the comparison theorem holds for k_-. We will now use Proposition C.10 to show that if the comparison theorem holds for $k-1$, then it holds for k also. Then the comparison theorem holds for every $k \geq k_-$ by induction, so the conclusion of Theorem 6.12 holds whenever f is a local function. The extension to quasilocal f follows readily by localization as in the proof of Lemma C.3.

We now complete the induction step. When the comparison theorem holds for $k-1$ (the induction hypothesis), we can apply Proposition C.10 with

$$\beta_i = \sum_{j\in I_{\leq k-1}} D_{ij}\, W^{-1}_{jj} a_j.$$

This gives

$$|\rho_{\leq k} f - \tilde{\rho}_{\leq k} f| \leq \sum_{j,q \in I_{\leq k-1}} \sum_{i,l \in I_k} \operatorname{osc}_i f \, D_{il} \, (W^{-1}R)_{lq} \, D_{qj} \, W_{jj}^{-1} a_j + \sum_{i,j \in I_{\leq k-1}} \operatorname{osc}_i f \, D_{ij} \, W_{jj}^{-1} a_j + \sum_{i,j \in I_k} \operatorname{osc}_i f \, D_{ij} \, W_{jj}^{-1} a_j$$

for every bounded measurable quasilocal function f on $\mathbb{S}_{\leq k}$ so that $\operatorname{osc}_i f < \infty \; \forall i$. To complete the proof, it therefore suffices to show that we have

$$D_{ij} = \sum_{q \in I_{\leq k-1}} \sum_{l \in I_k} D_{il} \, (W^{-1}R)_{lq} \, D_{qj} \qquad \text{for } i \in I_k, \; j \in I_{\leq k-1}.$$

To see this, note that as $R_{ij} = 0$ for $\tau(i) < \tau(j)$, we can write

$$\begin{aligned} D_{ij} &= \sum_{p=1}^{\infty} \sum_{\substack{j_1, \ldots, j_{p-1} \in I: \\ \tau(j) \leq \tau(j_1) \leq \cdots \leq \tau(j_{p-1}) \leq k}} (W^{-1}R)_{ij_{p-1}} \cdots (W^{-1}R)_{j_2 j_1} (W^{-1}R)_{j_1 j} \\ &= \sum_{p=1}^{\infty} \sum_{n=1}^{p} \sum_{l \in I_k} \sum_{q \in I_{\leq k-1}} (W^{-1}R)_{il}^{n-1} (W^{-1}R)_{lq} (W^{-1}R)_{qj}^{p-n} \end{aligned}$$

for $i \in I_k$ and $j \in I_{\leq k-1}$, where we have used that whenever $\tau(j_1) \leq \cdots \leq \tau(j_{p-1}) \leq k$ there exists $1 \leq n \leq p$ such that $j_1, \ldots, j_{p-n} \in I_{\leq k-1}$ and $j_{p-n+1}, \ldots, j_{p-1} \in I_k$. Rearranging the last expression yields the desired identity for D_{ij}, completing the proof for the case $k_- > -\infty$ (note that in this case the additional assumption (6.2) was not needed).

We now turn to the case that $k_- = -\infty$. Let us say that $(\beta_i)_{i \in I_{\leq k}}$ is a k-estimate if

$$|\rho_{\leq k} g - \tilde{\rho}_{\leq k} g| \leq \sum_{i \in I_{\leq k}} \operatorname{osc}_i g \, \beta_i$$

for every bounded measurable quasilocal function g on $\mathbb{S}_{\leq k}$ such that $\operatorname{osc}_i g < \infty \; \forall i$. Then the conclusion of Proposition C.10 can be reformulated as follows: if $(\beta_i)_{i \in I_{\leq k-1}}$ is a $(k-1)$-estimate, then $(\beta'_i)_{i \in I_{\leq k}}$ is a k-estimate with $\beta'_i = \beta_i$ for $i \in I_{\leq k-1}$ and

$$\beta'_i = \sum_{j \in I_{\leq k-1}} \sum_{l \in I_k} D_{il} \, (W^{-1}R)_{lj} \, \beta_j + \sum_{j \in I_k} D_{ij} \, W_{jj}^{-1} a_j$$

for $i \in I_k$. We can therefore repeatedly apply Proposition C.10 to extend an initial estimate. In particular, if we fix $k \in \mathbb{Z}$ and $n \geq 1$, and if $(\beta_i)_{i \in I_{\leq k-n}}$ is a $(k-n)$-estimate, then we can obtain a k-estimate $(\beta'_i)_{i \in I_{\leq k}}$ by iterating Proposition C.10 n times. We claim that

$$\beta'_i = \sum_{s=k-n+1}^{k-r} \left\{ \sum_{j \in I_{\leq k-n}} \sum_{l \in I_s} D_{il} \, (W^{-1}R)_{lj} \, \beta_j + \sum_{j \in I_s} D_{ij} \, W_{jj}^{-1} a_j \right\}$$

for $0 \le r \le n-1$ and $i \in I_{k-r}$. To see this, we proceed again by induction. As $(\beta_i)_{i\in I_{\le k-n}}$ is a $(k-n)$-estimate, the expression is valid for $r = n-1$ by Proposition C.10. Now suppose the expression is valid for all $u < r \le n-1$. Then we obtain

$$\begin{aligned}\beta'_i = &\sum_{j\in I_{\le k-n}} \sum_{l\in I_{k-u}} D_{il}\,(W^{-1}R)_{lj}\,\beta_j + \sum_{j\in I_{k-u}} D_{ij}\,W_{jj}^{-1}a_j \\ &+ \sum_{s=k-n+1}^{k-u-1} \sum_{j\in I_s} \sum_{l\in I_{k-u}} \sum_{t=k-n+1}^{s} \sum_{q\in I_{\le k-n}} \sum_{p\in I_t} D_{il}\,(W^{-1}R)_{lj}\,D_{jp}\,(W^{-1}R)_{pq}\,\beta_q \\ &+ \sum_{s=k-n+1}^{k-u-1} \sum_{j\in I_s} \sum_{l\in I_{k-u}} \sum_{t=k-n+1}^{s} \sum_{q\in I_t} D_{il}\,(W^{-1}R)_{lj}\,D_{jq}\,W_{qq}^{-1}a_q\end{aligned}$$

for $i \in I_{k-u}$ by Proposition C.10. Rearranging the sums yields

$$\begin{aligned}\beta'_i = &\sum_{j\in I_{\le k-n}} \sum_{l\in I_{k-u}} D_{il}\,(W^{-1}R)_{lj}\,\beta_j + \sum_{j\in I_{k-u}} D_{ij}\,W_{jj}^{-1}a_j \\ &+ \sum_{t=k-n+1}^{k-u-1} \Bigg\{ \sum_{q\in I_{\le k-n}} \sum_{p\in I_t} \bar{D}_{ip}\,(W^{-1}R)_{pq}\,\beta_q + \sum_{p\in I_t} \bar{D}_{ip}\,W_{pp}^{-1}a_p \Bigg\},\end{aligned}$$

for $i \in I_{k-u}$, where we have defined

$$\bar{D}_{ij} := \sum_{\ell=s}^{t-1} \sum_{q\in I_\ell} \sum_{l\in I_t} D_{il}\,(W^{-1}R)_{lq}\,D_{qj}$$

whenever $i \in I_t$ and $j \in I_s$ for $s < t$. But as $D_{qj} = 0$ when $\tau(q) < \tau(j)$, we have

$$\bar{D}_{ij} = \sum_{q\in I_{\le t-1}} \sum_{l\in I_t} D_{il}\,(W^{-1}R)_{lq}\,D_{qj} = D_{ij} \qquad \text{for } i \in I_t,\ j \in I_{\le t-1}$$

using the identity used in the proof for the case $k_- > -\infty$, and the claim follows.

We can now complete the proof for the case $k_- = -\infty$. It suffices to prove the theorem for a given local function f (the extension to quasilocal f follows readily as in the proof of Lemma C.3). Let us therefore fix a K-local function f for some $K \in \mathcal{I}$, and let $k = \max_{i\in K} \tau(i)$ and $n \ge 1$. By Lemma C.3, we find that $(\beta_i)_{i\in I_{\le k-n}}$ is trivially a $(k-n)$-estimate if we set $\beta_i = (\rho\otimes\tilde{\rho})\eta_i$ for $i \in I_{\le k-n}$. We therefore obtain

$$|\rho f - \tilde{\rho} f| \le \sum_{i,j\in I} \mathrm{osc}_i f\; D_{ij}\,W_{jj}^{-1}a_j + \sum_{i\in I} \sum_{j\in I_{\le k-n}} \mathrm{osc}_i f\; D_{ij}\,(\rho\otimes\tilde{\rho})\eta_j$$

from the k-estimate $(\beta'_i)_{i\in I_{\le k}}$ derived above, where we have used that $DW^{-1}R \le D$. But as f is local and $\mathrm{osc}_i f < \infty$ for all i by assumption, the second term vanishes as $n \to \infty$ by assumption (6.2). This completes the proof for the case $k_- = -\infty$. □

C.6 Block particle filter, improved analysis

In the remaining of this appendix we provide the proof of Theorem 6.13. We assume to work in the same setting introduced in Chapter 4. Recall the following three recursions:

$$\pi_n^\mu := \mathsf{F}_n \cdots \mathsf{F}_1 \mu, \qquad \tilde{\pi}_n^\mu := \tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_1 \mu, \qquad \hat{\pi}_n^\mu := \hat{\mathsf{F}}_n \cdots \hat{\mathsf{F}}_1 \mu,$$

where $\mathsf{F}_n := \mathsf{C}_n\mathsf{P}$, $\tilde{\mathsf{F}}_n := \mathsf{C}_n\mathsf{B}\mathsf{P}$, and $\hat{\mathsf{F}}_n := \mathsf{C}_n\mathsf{B}\mathsf{S}^N\mathsf{P}$. This allows to decompose the approximation error into two terms, one due to localization and one due to sampling

$$|||\pi_n^\mu - \hat{\pi}_n^\mu|||_J \le \underbrace{|||\pi_n^\mu - \tilde{\pi}_n^\mu|||_J}_{\text{bias}} + \underbrace{|||\tilde{\pi}_n^\mu - \hat{\pi}_n^\mu|||_J}_{\text{variance}}$$

by the triangle inequality (see Section 4.5.1). In the proof of Theorem 6.13, each of the terms on the right will be considered separately. The first term, which quantifies the bias due to the localization, will be bounded in Section C.6.1. The second term, which quantifies the sampling variance, will be bounded in Section C.6.2. Combining these two bounds completes the proof.

C.6.1 Bounding the bias

The goal of this section is to bound the bias term $\|\pi_n^\sigma - \tilde{\pi}_n^\sigma\|_J$, where we recall the definition

$$\|\mu - \nu\|_J := \sup_{f \in \mathcal{X}^J : |f| \le 1} |\mu f - \nu f|$$

the local total variation distance on the set of sites J. [Note that $\|\mu - \nu\|_J \le K$ for some $K \in \mathbb{R}$ evidently implies $|||\mu - \nu|||_J \le K$; the random measure norm $|||\cdot|||_J$ will be essential to bound the sampling error, but is irrelevant for the bias term.]

Let us first give an informal outline of the ideas behind the proof of the bias bound. While the filter π_n^σ is itself a high-dimensional distribution (defined on the set of sites V), we do not know how to obtain a tractable local update rule for it. We therefore cannot apply Theorem 6.4 directly. Instead, we will consider the *smoothing* distribution

$$\rho = \mathbf{P}^\sigma(X_1, \ldots, X_n \in \cdot\, | Y_1, \ldots, Y_n),$$

defined on the extended set of sites $I = \{1, \ldots, n\} \times V$ and configuration space $\mathbb{S} = \mathbb{X}^n$. As $(X_k^v, Y_k^v)_{(k,v)\in I}$ is a Markov random field (cf. Figure 4.1), we can read off a local update rule for ρ from the model definition. At the same time, as $\pi_n^\sigma = \mathbf{P}^\sigma(X_n \in \cdot\, | Y_1, \ldots, Y_n)$ is a marginal of ρ, we immediately obtain estimates for π_n^σ from estimates for ρ.

This basic idea relies on the probabilistic definition of the filter as a conditional distribution of a Markov random field: the filtering recursion (which was only introduced for computational purposes) plays no role in the analysis. The block filter $\tilde{\pi}_n^\sigma$, on the other hand, is *defined* in terms of a recursion and does not have an

intrinsic probabilistic interpretation. In order to handle the block filter, we will artificially cook up a probability measure $\tilde{\mathbf{P}}$ on $\mathbb{S}$ such that the block filter satisfies $\tilde{\pi}_n^\sigma = \tilde{\mathbf{P}}(X_n \in \cdot\,|Y_1, \ldots, Y_n)$, and set

$$\tilde{\rho} = \tilde{\mathbf{P}}(X_1, \ldots, X_n \in \cdot\,|Y_1, \ldots, Y_n).$$

This implies in particular that

$$\|\pi_n^\sigma - \tilde{\pi}_n^\sigma\|_J = \|\rho - \tilde{\rho}\|_{\{n\}\times J},$$

and we can now bound the bias term by applying Theorem 6.4.

To apply the comparison theorem we must choose a good cover $\mathcal{J}$. It is here that the full flexibility of Theorem 6.4, as opposed to the classical comparison theorem, comes into play. If we were to apply Theorem 6.4 with the singleton cover $\mathcal{J}_{\mathrm{s}} = \{\{i\} : i \in I\}$, we would recover the result of Theorem 4.2: in this case both the spatial and temporal interactions must be weak in order to ensure that $D = \sum_n (W^{-1}R)^n < \infty$. To avoid this problem, we work instead with larger blocks in the temporal direction. That is, our blocks $J \in \mathcal{J}$ will have the form $J = \{k+1, \ldots, k+q\} \times \{v\}$ for an appropriate choice of the block length q. The local update γ_x^J now behaves as q time steps of an ergodic Markov chain in $\mathbb{X}^v$: the temporal interactions decay geometrically with q, and can therefore be made arbitrarily small even if the interaction in one time step is arbitrarily strong. On the other hand, when we increase q there will be more nonzero terms in the matrix $W^{-1}R$. We must therefore ultimately tune the block length q appropriately to obtain the result of Theorem 6.13.

Remark C.12. *The approach used here to bound the bias directly using the comparison theorem is different than the one used in Chapter 4, which exploits the recursive property of the filter. The latter approach has a broader scope, as it does not rely on the ability to express the approximate filter as the marginal of a random field as we do above: this could be essential for the analysis of more sophisticated algorithms that do not admit such a representation. For the purposes of the current analysis, however, the present approach provides an alternative and somewhat shorter proof that is well adapted to the analysis of block particle filters.*

Remark C.13. *The problem under investigation is based on an interacting Markov chain model, and is therefore certainly dynamical in nature. Nonetheless, our proofs use Theorem 6.4 and not the one-sided Theorem 6.12. If we were to approximate the dynamics of the Markov chain X_n itself, it would be much more convenient to apply Theorem 6.12 as the model is already defined in terms of one-sided conditional distributions $p(x,z)\,\psi(dz)$. Unfortunately, when we condition on the observations Y_n, the one-sided conditional distributions take a complicated form that incorporates all the information in the future observations, whereas conditioning on all variables outside a block $J \in \mathcal{J}$ gives rise to relatively tractable expressions. For this reason, the static "space-time" picture remains the most convenient approach for the investigation of high-dimensional filtering problems.*

We now turn to the details of the proof. We first state the main result of this section.

Theorem C.14 (Bias term)**.** *Suppose there exist* $0 < \varepsilon, \delta < 1$ *such that*

$$\varepsilon q^v(x^v, z^v) \le p^v(x, z^v) \le \varepsilon^{-1} q^v(x^v, z^v),$$
$$\delta \le q^v(x^v, z^v) \le \delta^{-1}$$

for every $v \in V$ *and* $x, z \in \mathbb{X}$*, where* $q^v : \mathbb{X}^v \times \mathbb{X}^v \to \mathbb{R}_+$ *is a transition density with respect to* ψ^v*. Suppose also that we can choose* $q \in \mathbb{N}$ *and* $\beta > 0$ *such that*

$$c := 3q\Delta^2 e^{\beta(q+2r)}(1 - \varepsilon^{2(\Delta+1)}) + e^{\beta}(1 - \varepsilon^2\delta^2) + e^{\beta q}(1 - \varepsilon^2\delta^2)^q < 1.$$

Then we have

$$\|\pi_n^\sigma - \tilde{\pi}_n^\sigma\|_J \le \frac{2e^{\beta r}}{1 - c}\,(1 - \varepsilon^{2(q+1)\Delta})\ \operatorname{card} J\, e^{-\beta d(J, \partial K)}$$

for every $n \ge 0$*,* $\sigma \in \mathbb{X}$*,* $K \in \mathcal{K}$ *and* $J \subseteq K$*.*

In order to use the comparison theorem, we must have a method to construct couplings. Before we proceed to the proof of Theorem C.14, we begin by formulating two elementary results that will provide us with the necessary tools for this purpose.

Lemma C.15. *If probability measures* μ, ν, γ *satisfy* $\mu(A) \ge \alpha\gamma(A)$ *and* $\nu(A) \ge \alpha\gamma(A)$ *for every measurable set* A*, then there is a coupling* Q *of* μ, ν *such that* $\int \mathbf{1}_{x \ne z}\, Q(dx, dz) \le 1 - \alpha$*.*

Proof. Define $\tilde{\mu} = (\mu - \alpha\gamma)/(1 - \alpha)$, $\tilde{\nu} = (\nu - \alpha\gamma)/(1 - \alpha)$, and let

$$Qf = \alpha \int f(x, x)\, \gamma(dx) + (1 - \alpha) \int f(x, z)\, \tilde{\mu}(dx)\, \tilde{\nu}(dz).$$

The claim follows readily. □

Lemma C.16. *Let* $P_1, \ldots, P_q$ *be transition kernels on a measurable space* $\mathbb{T}$*, and define*

$$\mu_x(d\omega_1, \ldots, d\omega_q) = P_1(x, d\omega_1) P_2(\omega_1, d\omega_2) \cdots P_q(\omega_{q-1}, d\omega_q).$$

Suppose that there exist probability measures $\nu_1, \ldots, \nu_q$ *on* $\mathbb{T}$ *such that* $P_i(x, A) \ge \alpha\nu_i(A)$ *for every measurable set* A*,* $x \in \mathbb{T}$*, and* $i \le q$*. Then there exists for every* $x, z \in \mathbb{T}$ *a coupling* $Q_{x,z}$ *of* μ_x *and* μ_z *such that* $\int \mathbf{1}_{\omega_i \ne \omega_i'}\, Q_{x,z}(d\omega, d\omega') \le (1 - \alpha)^i$ *for every* $i \le q$*.*

Proof. Define the transition kernels $\tilde{P}_i = (P_i - \alpha\nu_i)/(1 - \alpha)$ and

$$\tilde{Q}_i f(x, z) = \alpha \int f(x', x')\, \nu_i(dx') + (1 - \alpha)\, \mathbf{1}_{x \ne z} \int f(x', z')\, \tilde{P}_i(x, dx')\, \tilde{P}_i(z, dz')$$
$$+ (1 - \alpha)\, \mathbf{1}_{x = z} \int f(x', x')\, \tilde{P}_i(x, dx').$$

Then $\tilde{Q}_i(x, z, \cdot\,)$ is a coupling of $P_i(x, \cdot\,)$ and $P_i(z, \cdot\,)$. Now define

$$Q_{x,z}(d\omega_1, d\omega_1', \ldots, d\omega_q, d\omega_q') = \tilde{Q}_1(x, z, d\omega_1, d\omega_1') \cdots \tilde{Q}_q(\omega_{q-1}, \omega_{q-1}', d\omega_q, d\omega_q').$$

The result follows readily once we note that $\int \mathbf{1}_{x' \ne z'}\, \tilde{Q}_i(x, z, dx', dz') \le (1 - \alpha)\, \mathbf{1}_{x \ne z}$. □

We can now proceed to the proof of Theorem C.14.

Proof of Theorem C.14. We begin by constructing a measure $\tilde{\mathbf{P}}$ that allows to describe the block filter $\tilde{\pi}_n^\sigma$ as a conditional distribution, as explained above. We fix the initial condition $\sigma \in \mathbb{X}$ throughout the proof (the dependence of various quantities on σ is implicit).

To construct $\tilde{\mathbf{P}}$, define for $K \in \mathcal{K}$ and $n \geq 1$ the function

$$h_n^K(x, z^{\partial K}) := \int \tilde{\pi}_{n-1}^\sigma(d\omega) \prod_{v \in \partial K} p^v(x^K \omega^{V \setminus K}, z^v).$$

Evidently h_n^K is a transition density with respect to $\bigotimes_{v \in \partial K} \psi^v$. Let

$$\tilde{p}_n(x, z) := \prod_{K \in \mathcal{K}} h_n^K(x, z^{\partial K}) \prod_{v \in K \setminus \partial K} p^v(x, z^v),$$

and define $\tilde{\mathsf{P}}_n \mu(dx') := \psi(dx') \int \tilde{p}_n(x, x')\, \mu(dx)$. Then $\tilde{\mathsf{P}}_n \tilde{\pi}_{n-1}^\sigma = \mathsf{BP} \tilde{\pi}_{n-1}^\sigma$ by construction for every $n \geq 1$, as $\tilde{\pi}_{n-1}^\sigma$ is a product measure across blocks. Thus we have

$$\pi_n^\sigma = \mathsf{C}_n \mathsf{P} \cdots \mathsf{C}_1 \mathsf{P} \delta_\sigma, \qquad \tilde{\pi}_n^\sigma = \mathsf{C}_n \tilde{\mathsf{P}}_n \cdots \mathsf{C}_1 \tilde{\mathsf{P}}_1 \delta_\sigma.$$

In particular, the filter and the block filter satisfy the same recursion with different transition densities p and $\tilde{p}_n$. We can therefore interpret the block filter as the filter corresponding to a time-inhomogeneous Markov chain with transition densities $\tilde{p}_n$: that is, if we set

$$\tilde{\mathbf{P}}[(X_1, \ldots, X_n, Y_1, \ldots, Y_n) \in A] := \int \mathbf{1}_A(x_1, \ldots, x_n, y_1, \ldots, y_n)\, \tilde{p}_1(\sigma, x_1) \prod_{k=2}^n \tilde{p}_k(x_{k-1}, x_k)\, g(x_k, y_k)\, \psi(dx_k)\, \varphi(dy_k)$$

(note that $\mathbf{P}^\sigma$ satisfies the same formula where $\tilde{p}_k$ is replaced by p), we can write

$$\tilde{\pi}_n^\sigma = \tilde{\mathbf{P}}(X_n \in \cdot\, | Y_1, \ldots, Y_n).$$

Let us emphasize that the transition densities $\tilde{p}_k$ and operators $\tilde{\mathsf{P}}_k$ themselves depend on the initial condition σ, which is certainly not the case for the regular filter. However, since σ is fixed throughout the proof, this is irrelevant for our computations.

From now on we fix $n \geq 1$ in the remainder of the proof. Let

$$\rho = \mathbf{P}^\sigma(X_1, \ldots, X_n \in \cdot\, | Y_1, \ldots, Y_n), \qquad \tilde{\rho} = \tilde{\mathbf{P}}(X_1, \ldots, X_n \in \cdot\, | Y_1, \ldots, Y_n).$$

Then ρ and $\tilde{\rho}$ are probability measures on $\mathbb{S} = \mathbb{X}^n$, which is naturally indexed by the set of sites $I = \{1, \ldots, n\} \times V$ (the observation sequence on which we condition is arbitrary and can be considered fixed throughout the proof). The proof now proceeds by applying Theorem 6.4 to $\rho, \tilde{\rho}$, the main difficulty being the construction of a coupled update rule.

Fix $q \geq 1$. We first specify the cover $\mathcal{J} = \{J_l^v : 1 \leq l \leq \lceil n/q \rceil, v \in V\}$ as follows:

$$J_l^v := \{(l-1)q+1, \ldots, lq \wedge n\} \times \{v\} \qquad \text{for } 1 \leq l \leq \lceil n/q \rceil,\ v \in V.$$

We choose the natural local updates $\gamma_x^J(dz^J) = \rho(dz^J|x^{I\setminus J})$ and $\tilde{\gamma}_x^J(dz^J) = \tilde{\rho}(dz^J|x^{I\setminus J})$, and postpone the construction of the coupled updates $Q_{x,z}^J$ and $\hat{Q}_x^J$ to be done below. Now note that the cover $\mathcal{J}$ is in fact a partition of I; thus Theorem 6.4 yields

$$\|\pi_n^\sigma - \tilde{\pi}_n^\sigma\|_J = \|\rho - \tilde{\rho}\|_{\{n\}\times J} \leq 2 \sum_{i \in \{n\}\times J} \sum_{j \in I} D_{ij}\, b_j$$

provided that $D = \sum_{k=0}^\infty C^k < \infty$ (cf. Corollary 6.8), where

$$C_{ij} = \sup_{\substack{x,z \in \mathbb{S}: \\ x^{I\setminus\{j\}} = z^{I\setminus\{j\}}}} \int \mathbf{1}_{\omega_i \neq \omega_i'}\, Q_{x,z}^{J(i)}(d\omega, d\omega'), \qquad b_i = \sup_{x \in \mathbb{S}} \int \mathbf{1}_{\omega_i \neq \omega_i'}\, \hat{Q}_x^{J(i)}(d\omega, d\omega'),$$

and where we write $J(i)$ for the unique block $J \in \mathcal{J}$ that contains $i \in I$. To put this bound to good use, we must introduce coupled updates $Q_{x,z}^J$ and $\hat{Q}_x^J$ and estimate C_{ij} and b_j.

Let us fix until further notice a block $J = J_l^v \in \mathcal{J}$. We will consider first the case that $1 < l < \lceil n/q \rceil$; the cases $l = 1, \lceil n/q \rceil$ will follow subsequently using the identical proof. Let $s = (l-1)q$. Then we can compute explicitly the local update rule

$$\gamma_x^J(A) = \frac{\int \mathbf{1}_A(x^J)\, p^v(x_s, x_{s+1}^v) \prod_{m=s+1}^{s+q} g^v(x_m^v, Y_m^v) \prod_{w \in N(v)} p^w(x_m, x_{m+1}^w)\, \psi^v(dx_m^v)}{\int p^v(x_s, x_{s+1}^v) \prod_{m=s+1}^{s+q} g^v(x_m^v, Y_m^v) \prod_{w \in N(v)} p^w(x_m, x_{m+1}^w)\, \psi^v(dx_m^v)}$$

using Bayes' formula, the definition of $\mathbf{P}^\sigma$ (in the same form as the above definition of $\tilde{\mathbf{P}}$), and that $p^v(x, z^v)$ depends only on $x^{N(v)}$. We now construct couplings $Q_{x,z}^J$ of γ_x^J and γ_z^J where x, z differ only at the site $j = (k, w) \in I$. We distinguish the following cases:

1. $k = s$, $w \in N(v)\setminus\{v\}$;
2. $k = s$, $w = v$;
3. $k \in \{s+1, \ldots, s+q\}$, $w \in \bigcup_{u \in N(v)} N(u)\setminus\{v\}$;
4. $k = s+q+1$, $w \in N(v)\setminus\{v\}$;
5. $k = s+q+1$, $w = v$.

It is easily verified by inspection that γ_x^J does not depend on x_k^w except in one of the above cases. Thus when j satisfies none of the above conditions, we can set $C_{ij} = 0$ for $i \in J$.

Case 1. Note that

$$\gamma_x^J(A) \geq \varepsilon^2 \frac{\int \mathbf{1}_A(x^J)\, q^v(x_s^v, x_{s+1}^v) \prod_{m=s+1}^{s+q} g^v(x_m^v, Y_m^v) \prod_{w\in N(v)} p^w(x_m, x_{m+1}^w)\, \psi^v(dx_m^v)}{\int q^v(x_s^v, x_{s+1}^v) \prod_{m=s+1}^{s+q} g^v(x_m^v, Y_m^v) \prod_{w\in N(v)} p^w(x_m, x_{m+1}^w)\, \psi^v(dx_m^v)},$$

and the right hand side does not depend on x_s^w for $w \neq v$. Thus whenever $x, z \in \mathbb{S}$ satisfy $x^{I\setminus\{j\}} = z^{I\setminus\{j\}}$ for $j = (s, w)$ with $w \in N(v)\setminus\{v\}$, we can construct a coupling $Q_{x,z}^J$ using Lemma C.15 such that $C_{ij} \leq 1 - \varepsilon^2$ for every $i \in J$.

Case 2. Define the transition kernels on $\mathbb{X}^v$

$$P_{k,x}(\omega, A) = \frac{\int \mathbf{1}_A(x_k^v)\, p^v(\omega x_{k-1}^{V\setminus\{v\}}, x_k^v) \prod_{m=k}^{s+q} g^v(x_m^v, Y_m^v) \prod_{w\in N(v)} p^w(x_m, x_{m+1}^w)\, \psi^v(dx_m^v)}{\int p^v(\omega x_{k-1}^{V\setminus\{v\}}, x_k^v) \prod_{m=k}^{s+q} g^v(x_m^v, Y_m^v) \prod_{w\in N(v)} p^w(x_m, x_{m+1}^w)\, \psi^v(dx_m^v)}$$

for $k = s+1, \ldots, s+q$. By construction, $P_{k,x}(x_{k-1}^v, dx_k^v) = \gamma_x^J(dx_k^v | x_{s+1}^v, \ldots, x_{k-1}^v)$, so we are in the setting of Lemma C.16. Moreover, we can estimate

$$P_{k,x}(\omega, A) \geq \varepsilon^2\delta^2 \frac{\int \mathbf{1}_A(x_k^v) \prod_{m=k}^{s+q} g^v(x_m^v, Y_m^v) \prod_{w\in N(v)} p^w(x_m, x_{m+1}^w)\, \psi^v(dx_m^v)}{\int \prod_{m=k}^{s+q} g^v(x_m^v, Y_m^v) \prod_{w\in N(v)} p^w(x_m, x_{m+1}^w)\, \psi^v(dx_m^v)},$$

where the right hand side does not depend on ω. Thus whenever $x, z \in \mathbb{S}$ satisfy $x^{I\setminus\{j\}} = z^{I\setminus\{j\}}$ for $j = (s, v)$, we can construct a coupling $Q_{x,z}^J$ using Lemma C.16 such that $C_{ij} \leq (1 - \varepsilon^2\delta^2)^{k-s}$ for $i = (k, v)$ with $k = s+1, \ldots, s+q$.

Case 3. Fix $k \in \{s+1, \ldots, s+q\}$ and $u \neq v$. Note that

$$\gamma_x^J(A) \geq \varepsilon^{2(\Delta+1)} \times \frac{\int \mathbf{1}_A(x^J)\, p^v(x_s, x_{s+1}^v) \prod_{m=s+1}^{s+q} g^v(x_m^v, Y_m^v) \prod_{w\in N(v)} \beta_m^w(x_m, x_{m+1}^w)\, \psi^v(dx_m^v)}{\int p^v(x_s, x_{s+1}^v) \prod_{m=s+1}^{s+q} g^v(x_m^v, Y_m^v) \prod_{w\in N(v)} \beta_m^w(x_m, x_{m+1}^w)\, \psi^v(dx_m^v)}$$

where we set $\beta_m^w(x_m, x_{m+1}^w) = q^w(x_m^w, x_{m+1}^w)$ if either $m = k$ or $m = k-1$ and $w = u$, and $\beta_m^w(x_m, x_{m+1}^w) = p^w(x_m, x_{m+1}^w)$ otherwise. The right hand side of this expression does not depend on x_k^u as the terms $q^w(x_m^w, x_{m+1}^w)$ for $w \neq v$ cancel in the numerator and denominator. Thus whenever $x, z \in \mathbb{S}$ satisfy $x^{I\setminus\{j\}} = z^{I\setminus\{j\}}$ for $j = (k, u)$, we can construct a coupling $Q_{x,z}^J$ using Lemma C.15 such that $C_{ij} \leq 1 - \varepsilon^{2(\Delta+1)}$ for every $i \in J$.

Case 4. Let $u \in N(v)\setminus v$. Note that

$$\gamma_x^J(A) \geq \varepsilon^2 \frac{\int \mathbf{1}_A(x^J)\, p^v(x_s^v, x_{s+1}^v) \prod_{m=s+1}^{s+q} g^v(x_m^v, Y_m^v) \prod_{w\in N(v)} \beta_m^w(x_m, x_{m+1}^w)\, \psi^v(dx_m^v)}{\int p^v(x_s^v, x_{s+1}^v) \prod_{m=s+1}^{s+q} g^v(x_m^v, Y_m^v) \prod_{w\in N(v)} \beta_m^w(x_m, x_{m+1}^w)\, \psi^v(dx_m^v)},$$

where we set $\beta_m^w(x_m, x_{m+1}^w) = q^w(x_m^w, x_{m+1}^w)$ if $m = s+q$ and $w = u$, and we set $\beta_m^w(x_m, x_{m+1}^w) = p^w(x_m, x_{m+1}^w)$ otherwise. The right hand side does not depend on

x^u_{s+q+1} as the term $q^u(x^u_{s+q}, x^u_{s+q+1})$ cancels in the numerator and denominator. Thus whenever $x, z \in \mathbb{S}$ satisfy $x^{I\setminus\{j\}} = z^{I\setminus\{j\}}$ for $j = (s+q+1, u)$, we can construct a coupling $Q^J_{x,z}$ using Lemma C.15 such that $C_{ij} \le 1 - \varepsilon^2$ for every $i \in J$.

Case 5. Define for $k = s+1, \dots, s+q$ the transition kernels on $\mathbb{X}^v$

$$P_{k,x}(\omega, A) = \frac{\int \mathbf{1}_A(x^v_k)\, p^v(x_s, x^v_{s+1}) \prod_{m=s+1}^k g^v(x^v_m, Y^v_m) \prod_{w\in N(v)} \beta^w_{m,\omega}(x_m, x^w_{m+1})\, \psi^v(dx^v_m)}{\int p^v(x_s, x^v_{s+1}) \prod_{m=s+1}^k g^v(x^v_m, Y^v_m) \prod_{w\in N(v)} \beta^w_{m,\omega}(x_m, x^w_{m+1})\, \psi^v(dx^v_m)},$$

where we set $\beta^w_{m,\omega}(x_m, x^w_{m+1}) = p^v(x_k, \omega)$ if $m = k$ and $w = v$, and $\beta^w_{m,\omega}(x_m, x^w_{m+1}) = p^w(x_m, x^w_{m+1})$ otherwise. By construction, $P_{k,x}(x^v_{k+1}, dx^v_k) = \gamma^J_x(dx^v_k | x^v_{k+1}, \dots, x^v_{s+q})$, so we are in the setting of Lemma C.16. Moreover, we can estimate

$$P_{k,x}(\omega, A) \ge \varepsilon^2\delta^2 \times \frac{\int \mathbf{1}_A(x^v_k)\, p^v(x_s, x^v_{s+1}) \prod_{m=s+1}^k g^v(x^v_m, Y^v_m) \prod_{w\in N(v)} \beta^w_{m}(x_m, x^w_{m+1})\, \psi^v(dx^v_m)}{\int p^v(x_s, x^v_{s+1}) \prod_{m=s+1}^k g^v(x^v_m, Y^v_m) \prod_{w\in N(v)} \beta^w_{m}(x_m, x^w_{m+1})\, \psi^v(dx^v_m)},$$

where $\beta^w_m(x_m, x^w_{m+1}) = 1$ if $m = k$ and $w = v$, and $\beta^w_m(x_m, x^w_{m+1}) = p^w(x_m, x^w_{m+1})$ otherwise. Note that the right hand side does not depend on ω. Thus whenever $x, z \in \mathbb{S}$ satisfy $x^{I\setminus\{j\}} = z^{I\setminus\{j\}}$ for $j = (s+q+1, v)$, we can construct a coupling $Q^J_{x,z}$ using Lemma C.16 such that $C_{ij} \le (1-\varepsilon^2\delta^2)^{s+q+1-k}$ for $i = (k, v)$ with $k = s+1, \dots, s+q$.

We have now constructed coupled updates $Q^J_{x,z}$ for every pair $x, z \in \mathbb{S}$ that differ only at one point. Collecting the above bounds on C_{ij}, we can estimate

$$\begin{aligned}
&\sum_{(k',v')\in I} e^{\beta\{|k-k'|+d(v,v')\}} C_{(k,v)(k',v')} \\
&\quad\le 2e^{\beta(q+r)}(1-\varepsilon^2)\Delta + e^{\beta(q+2r)}(1-\varepsilon^{2(\Delta+1)})\Delta^2 q \\
&\qquad + e^{\beta(k-s)}(1-\varepsilon^2\delta^2)^{k-s} + e^{\beta(s+q+1-k)}(1-\varepsilon^2\delta^2)^{s+q+1-k} \\
&\quad\le 3q\Delta^2 e^{\beta(q+2r)}(1-\varepsilon^{2(\Delta+1)}) + e^{\beta}(1-\varepsilon^2\delta^2) + e^{\beta q}(1-\varepsilon^2\delta^2)^q =: c
\end{aligned}$$

whenever $(k, v) \in J$. In the last line, we have used that $\alpha^{x+1} + \alpha^{q-x}$ is a convex function of $x \in [0, q-1]$, and therefore attains its maximum on the endpoints $x = 0, q-1$.

Up to this point we have considered an arbitrary block $J = J^v_l \in \mathcal{J}$ with $1 < l < \lceil n/q \rceil$. It is however evident that the identical proof holds for the boundary blocks $l = 1, \lceil n/q \rceil$, except that for $l = 1$ we only need to consider Cases 3–5 above and for $l = \lceil n/q \rceil$ we only need to consider Cases 1–3 above. As all the estimates are otherwise identical, the corresponding bounds on C_{ij} are at most as large as those in the case $1 < l < \lceil n/q \rceil$. Thus

$$\|C\|_{\infty,\beta m} := \max_{i\in I} \sum_{j\in I} e^{\beta m(i,j)} C_{ij} \le c,$$

where we define the metric $m(i,j) = |k-k'| + d(v,v')$ for $(k,v) \in I$ and $(k',v') \in I$.

Our next order of business is to construct couplings $\hat{Q}_x^J$ of γ_x^J and $\tilde{\gamma}_x^J$ and to estimate the coefficients b_i. To this end, let us first note that $h_n^K(x, z^{\partial K})$ depends only on $x^{\partial^2 K}$, where

$$\partial^2 K := \bigcup_{w \in \partial K} N(w) \cap K$$

is the subset of vertices in K that can interact with vertices outside K in two time steps. It is easily seen that $\gamma_x^J = \tilde{\gamma}_x^J$, and that we can therefore choose $b_i = 0$ for $i \in J$, unless $J = J_l^v$ with $v \in \partial^2 K$ for some $K \in \mathcal{K}$. In the latter case we obtain by Bayes' formula

$$\tilde{\gamma}_x^J(A) = \frac{\int \mathbf{1}_A(x^J) \prod_{m=s}^{s+q} g^v(x_m^v, Y_m^v)\, h_{m+1}^K(x_m, x_{m+1}^{\partial K}) \prod_{w \in N(v) \cap K \setminus \partial K} p^w(x_m, x_{m+1}^w)\, \psi(dx^J)}{\int \prod_{m=s}^{s+q} g^v(x_m^v, Y_m^v)\, h_{m+1}^K(x_m, x_{m+1}^{\partial K}) \prod_{w \in N(v) \cap K \setminus \partial K} p^w(x_m, x_{m+1}^w)\, \psi(dx^J)}$$

for $1 < l < \lceil n/q \rceil$, where $s = (l-1)q$ and $\psi(dx^J) = \bigotimes_{(k,v) \in J} \psi^v(dx_k^v)$. Note that

$$\prod_{w \in N(v) \setminus (K \setminus \partial K)} p^w(x, z^w) \geq \varepsilon^{\Delta} \prod_{w \in N(v) \setminus (K \setminus \partial K)} q^w(x^w, z^w),$$

while

$$h_m^K(x, z^{\partial K}) \geq \varepsilon^{\Delta} \prod_{w \in N(v) \cap \partial K} q^w(x^w, z^w) \int \tilde{\pi}_{m-1}^{\sigma}(d\omega) \prod_{w \in \partial K \setminus N(v)} p^w(x^K \omega^{V \setminus K}, z^w).$$

We can therefore estimate $\gamma_x^J(A) \geq \varepsilon^{2(q+1)\Delta} \Gamma(A)$ and $\tilde{\gamma}_x^J(A) \geq \varepsilon^{2(q+1)\Delta} \Gamma(A)$ with

$$\Gamma(A) = \frac{\int \mathbf{1}_A(x^J) \prod_{m=s}^{s+q} g^v(x_m^v, Y_m^v)\, \beta(x_m^v, x_{m+1}^v) \prod_{w \in N(v) \cap K \setminus \partial K} p^w(x_m, x_{m+1}^w)\, \psi(dx^J)}{\int \prod_{m=s}^{s+q} g^v(x_m^v, Y_m^v)\, \beta(x_m^v, x_{m+1}^v) \prod_{w \in N(v) \cap K \setminus \partial K} p^w(x_m, x_{m+1}^w)\, \psi(dx^J)},$$

where $\beta(x,z) = q^v(x,z)$ if $v \in \partial K$ and $\beta(x,z) = 1$ if $v \in \partial^2 K \setminus \partial K$. Thus we can construct a coupling $\hat{Q}_x^J$ using Lemma C.15 such that $b_i \leq 1 - \varepsilon^{2(q+1)\Delta}$ for all $i \in J$ in the case $1 < l < \lceil n/q \rceil$. The same conclusion follows for $l = 1, \lceil n/q \rceil$ by the identical proof.

We are now ready to put everything together. As $\|\cdot\|_{\infty,\beta m}$ is a matrix norm, we have

$$\|D\|_{\infty,\beta m} \leq \sum_{k=0}^{\infty} \|C\|_{\infty,\beta m}^k \leq \frac{1}{1-c} < \infty.$$

Thus $D < \infty$, to we can apply the comparison theorem. Moreover,

$$\sup_{i \in J} \sum_{j \in J'} D_{ij} = \sup_{i \in J} e^{-\beta m(i,J')} \sum_{j \in J'} e^{\beta m(i,J')} D_{ij} \leq e^{-\beta m(J,J')} \|D\|_{\infty,\beta m}.$$

Thus we obtain

$$\|\pi_n^\sigma - \tilde{\pi}_n^\sigma\|_J \le 2(1-\varepsilon^{2(q+1)\Delta}) \sum_{i\in\{n\}\times J} \sum_{j\in\{1,\dots,n\}\times\partial^2 K} D_{ij}$$
$$\le \frac{2}{1-c}(1-\varepsilon^{2(q+1)\Delta})\ \mathrm{card}\, J\, e^{-\beta d(J,\partial^2 K)}.$$

But clearly $d(J,\partial^2 K) \ge d(J,\partial K) - r$, and the proof is complete. □

Remark C.17. *In the proof of Theorem C.14 (and similarly for Theorem C.20 below), we apply the comparison theorem with a nonoverlapping cover* $\{(l-1)q+1,\dots,lq\wedge n\}$, $l \le \lceil n/q \rceil$. *Working instead with overlapping blocks* $\{s+1,\dots,s+q\}$, $s \le n-q$ *would give somewhat better estimates at the expense of even more tedious computations.*

C.6.2 Bounding the variance

We now turn to the problem of bounding the variance term $\|\!|\!|\tilde{\pi}_n^\sigma - \hat{\pi}_n^\sigma|\!|\!|_J$. We will follow the basic approach taken in Section 4.5.4 and Section A.6, where a detailed discussion of the requisite ideas can be found. In this section we develop the necessary changes to the proof in Section A.6.

At the heart of the proof of the variance bound lies a stability result for the block filter, Proposition A.13. This result must be modified in the present setting to account for the different assumptions on the spatial and temporal correlations. This will be done next, using the generalized comparison Theorem 6.4 much as in the proof of Theorem C.14.

Proposition C.18. *Suppose there exist* $0 < \varepsilon, \delta < 1$ *such that*

$$\varepsilon q^v(x^v, z^v) \le p^v(x, z^v) \le \varepsilon^{-1} q^v(x^v, z^v),$$
$$\delta \le q^v(x^v, z^v) \le \delta^{-1}$$

for every $v \in V$ *and* $x, z \in \mathbb{X}$, *where* $q^v : \mathbb{X}^v \times \mathbb{X}^v \to \mathbb{R}_+$ *is a transition density with respect to* ψ^v. *Suppose also that we can choose* $q \in \mathbb{N}$ *and* $\beta > 0$ *such that*

$$c := 3q\Delta^2 e^{\beta q}(1-\varepsilon^{2(\Delta+1)}) + e^{\beta}(1-\varepsilon^2\delta^2) + e^{\beta q}(1-\varepsilon^2\delta^2)^q < 1.$$

Then we have

$$\|\tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_{s+1}\delta_\sigma - \tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_{s+1}\delta_{\tilde{\sigma}}\|_J \le \frac{2}{1-c}\ \mathrm{card}\, J\, e^{-\beta(n-s)}$$

for every $s < n$, $\sigma, \tilde{\sigma} \in \mathbb{X}$, $K \in \mathcal{K}$ *and* $J \subseteq K$.

Proof. We fix throughout the proof $n > 0$, $K \in \mathcal{K}$, and $J \subseteq K$. We will also assume for notational simplicity that $s = 0$. As $\tilde{\mathsf{F}}_k$ differ for different k only by their dependence on different observations Y_k, and as the conclusion of the Proposition is independent of the observations, the conclusion for $s = 0$ extends trivially to any $s < n$.

As in Theorem C.14, the idea behind the proof is to introduce a Markov random field ρ of which the block filter is a marginal, followed by an application of the generalized comparison theorem. Unfortunately, the construction in the proof of Theorem C.14 is not appropriate in the present setting, as there all the local interactions depend on the initial condition σ. That was irrelevant in Theorem C.14 where the initial condition was fixed, but is fatal in the present setting where we aim to understand a perturbation to the initial condition. Instead, we will use a more elaborate construction of ρ introduced in Section A.6, called the *computation tree*. We begin by recalling this construction.

Define for $K' \in \mathcal{K}$ the block neighborhood $N(K') := \{K'' \in \mathcal{K} : d(K', K'') \le r\}$ (we recall that $\operatorname{card} N(K') \le \Delta_{\mathcal{K}}$). We can evidently write

$$\mathsf{B}^{K'}\tilde{\mathsf{F}}_s \bigotimes_{K'' \in \mathcal{K}} \mu^{K''} = \mathsf{C}_s^{K'}\mathsf{P}^{K'} \bigotimes_{K'' \in N(K')} \mu^{K''},$$

where we define for any probability η on $\mathbb{X}^{K'}$

$$(\mathsf{C}_s^{K'}\eta)(A) := \frac{\int \mathbf{1}_A(x^{K'}) \prod_{v \in K'} g^v(x^v, Y_s^v)\, \eta(dx^{K'})}{\int \prod_{v \in K'} g^v(x^v, Y_s^v)\, \eta(dx^{K'})},$$

and for any probability η on $\mathbb{X}^{\bigcup_{K'' \in N(K')} K''}$

$$(\mathsf{P}^{K'}\eta)(A) := \int \mathbf{1}_A(x^{K'}) \prod_{v \in K'} p^v(z, x^v)\, \psi^v(dx^v)\, \eta(dz).$$

Iterating this identity yields

$$\mathsf{B}^K\tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_1\delta_\sigma = \mathsf{C}_n^K\mathsf{P}^K \bigotimes_{K_{n-1} \in N(K)} \Big[\cdots \mathsf{C}_2^{K_2}\mathsf{P}^{K_2} \bigotimes_{K_1 \in N(K_2)} \Big[\mathsf{C}_1^{K_1}\mathsf{P}^{K_1} \bigotimes_{K_0 \in N(K_1)} \delta_{\sigma^{K_0}} \Big] \cdots \Big].$$

The nested products can be naturally viewed as defining a tree.

To formalize this idea, define the tree index set (we will write $K_n := K$ for simplicity)

$$T := \{[K_u \cdots K_{n-1}] : 0 \le u < n,\ K_s \in N(K_{s+1}) \text{ for } u \le s < n\} \cup \{[\varnothing]\}.$$

The root of the tree $[\varnothing]$ represents the block K at time n, while $[K_u \cdots K_{n-1}]$ represents a duplicate of the block K_u at time u that affects the root along the branch $K_u \to K_{u+1} \to \cdots \to K_{n-1} \to K$. The set of sites corresponding to the computation tree is

$$I = \{[K_u \cdots K_{n-1}]v : [K_u \cdots K_{n-1}] \in T,\ v \in K_u\} \cup \{[\varnothing]v : v \in K\},$$

and the corresponding configuration space is $\mathbb{S} = \prod_{i \in I} \mathbb{X}^i$ with $\mathbb{X}^{[t]v} := \mathbb{X}^v$. The following tree notation will be used throughout the proof. Define for vertices of the

tree T the depth $d([K_u \cdots K_{n-1}]) := u$ and $d([\varnothing]) := n$. For every site $[t]v \in I$, we define the associated vertex $v(i) := v$ and depth $d(i) := d([t])$. Define also the sets $I_+ := \{i \in I : d(i) > 0\}$ and $T_0 := \{[t] \in T : d([t]) = 0\}$ of non-leaf sites and leaf vertices, respectively. Define

$$c([K_u \cdots K_{n-1}]v) := \{[K_{u-1} \cdots K_{n-1}]v' : K_{u-1} \in N(K_u),\ v' \in N(v)\},$$

and similarly for $c([\varnothing]v)$: that is, $c(i)$ is the set of children of the site $i \in I$ in the computation tree. Finally, we will frequently identify a tree vertex $[K_u \cdots K_{n-1}] \in T$ with the corresponding set of sites $\{[K_u \cdots K_{n-1}]v : v \in K_u\}$, and analogously for $[\varnothing]$.

Having introduced the tree structure, we now define probability measures $\rho, \tilde{\rho}$ on $\mathbb{S}$ by

$$\rho(A) = \frac{\int \mathbf{1}_A(x) \prod_{i \in I_+} p^{v(i)}(x^{c(i)}, x^i)\, g^{v(i)}(x^i, Y^i)\, \psi^{v(i)}(dx^i) \prod_{[t] \in T_0} \delta_{\sigma^{[t]}}(dx^{[t]})}{\int \prod_{i \in I_+} p^{v(i)}(x^{c(i)}, x^i)\, g^{v(i)}(x^i, Y^i)\, \psi^{v(i)}(dx^i) \prod_{[t] \in T_0} \delta_{\sigma^{[t]}}(dx^{[t]})},$$
$$\tilde{\rho}(A) = \frac{\int \mathbf{1}_A(x) \prod_{i \in I_+} p^{v(i)}(x^{c(i)}, x^i)\, g^{v(i)}(x^i, Y^i)\, \psi^{v(i)}(dx^i) \prod_{[t] \in T_0} \delta_{\tilde{\sigma}^{[t]}}(dx^{[t]})}{\int \prod_{i \in I_+} p^{v(i)}(x^{c(i)}, x^i)\, g^{v(i)}(x^i, Y^i)\, \psi^{v(i)}(dx^i) \prod_{[t] \in T_0} \delta_{\tilde{\sigma}^{[t]}}(dx^{[t]})},$$

where we write $\sigma^{[K_0 \cdots K_{n-1}]} := \sigma^{K_0}$ and $Y^i := Y^{v(i)}_{d(i)}$ for simplicity. Then, by construction, the measure $\mathsf{B}^K \tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_1 \delta_\sigma$ coincides with the marginal of ρ on the root of the computation tree, while $\mathsf{B}^K \tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_1 \delta_{\tilde{\sigma}}$ coincides with the marginal of $\tilde{\rho}$ on the root of the tree: this is easily seen by expanding the above nested product identity. In particular, we obtain

$$\|\tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_1 \delta_\sigma - \tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_1 \delta_{\tilde{\sigma}}\|_J = \|\rho - \tilde{\rho}\|_{[\varnothing]J},$$

and we aim to apply the comparison theorem to estimate this quantity.

The construction of the computation tree that we have just given is identical to the construction in Section A.6. We deviate from the proof of Appendix A from this point onward, since we must use Theorem 6.4 instead of the classical Dobrushin comparison theorem to account for the distinction between temporal and spatial correlations in the present setting.

Fix $q \geq 1$. In analogy with the proof of Theorem C.14, we consider a cover $\mathcal{J}$ consisting of blocks of sites $i \in I$ such that $(l-1)q < d(i) \leq lq \wedge n$ and $v(i) = v$. In the present setting, however, the same vertex v is duplicated many times in the tree, so that we end up with many disconnected blocks of different lengths. To keep track of these blocks, define

$$I_0 := \{i \in I : d(i) = 0\}, \qquad I_l := \{i \in I : d(i) = (l-1)q + 1\}$$

for $1 \leq l \leq \lceil n/q \rceil$, and let

$$\ell([K_u, \ldots, K_{n-1}]v) := \max\{s \geq u : K_u = K_{u+1} = \cdots = K_s\}.$$

We now define the cover $\mathcal{J}$ as

$$\mathcal{J} = \{J^i_l : 0 \leq l \leq \lceil n/q \rceil,\ i \in I_l\},$$

where

$$J_0^i := \{i\}, \qquad J_l^i := \{[K_u \cdots K_{n-1}]v : (l-1)q+1 \le u \le lq \wedge \ell(i)\}$$

for $i = [K_{(l-1)q+1} \cdots K_{n-1}]v \in I_l$ and $1 \le l \le \lceil n/q \rceil$. It is easily seen that $\mathcal{J}$ is in fact a partition of of the computation tree I into linear segments.

Having defined the cover $\mathcal{J}$, we must now consider a suitable coupled update rule. We will choose the natural local updates $\gamma_x^J(dz^J) = \rho(dz^J|x^{I\setminus J})$ and $\tilde{\gamma}_x^J(dz^J) = \tilde{\rho}(dz^J|x^{I\setminus J})$, with the coupled updates $Q_{x,z}^J$ and $\hat{Q}_x^J$ to be constructed below. Then Theorem 6.4 yields

$$\|\tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_1 \delta_\sigma - \tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_1 \delta_{\tilde{\sigma}}\|_J = \|\rho - \tilde{\rho}\|_{[\varnothing]J} \le 2 \sum_{i \in [\varnothing]J} \sum_{j \in I} D_{ij}\, b_j$$

provided that $D = \sum_{k=0}^\infty C^k < \infty$ (cf. Corollary 6.8), where

$$C_{ij} = \sup_{\substack{x,z \in \mathbb{S}: \\ x^{I\setminus\{j\}} = z^{I\setminus\{j\}}}} \int \mathbf{1}_{\omega_i \neq \omega_i'}\, Q_{x,z}^{J(i)}(d\omega, d\omega'), \qquad b_i = \sup_{x \in \mathbb{S}} \int \mathbf{1}_{\omega_i \neq \omega_i'}\, \hat{Q}_x^{J(i)}(d\omega, d\omega'),$$

and where we write $J(i)$ for the unique block $J \in \mathcal{J}$ that contains $i \in I$. To put this bound to good use, we must introduce coupled updates $Q_{x,z}^J$ and $\hat{Q}_x^J$ and estimate C_{ij} and b_j.

Let us fix until further notice a block $J = J_l^i \in \mathcal{J}$ with $i = [K_{(l-1)q+1} \cdots K_{n-1}]v \in I_l$ and $1 \le l \le \lceil n/q \rceil$. From the definition of ρ, we can compute explicitly

$$\gamma_x^J(A) = \frac{\int \mathbf{1}_A(x^J)\, p^v(x^{c(i)}, x^i) \prod_{a \in I_+ : J \cap c(a) \neq \varnothing} p^{v(a)}(x^{c(a)}, x^a) \prod_{b \in J} g^v(x^b, Y^b)\, \psi^v(dx^b)}{\int p^v(x^{c(i)}, x^i) \prod_{a \in I_+ : J \cap c(a) \neq \varnothing} p^{v(a)}(x^{c(a)}, x^a) \prod_{b \in J} g^v(x^b, Y^b)\, \psi^v(dx^b)}$$

using the Bayes formula. We now proceed to construct couplings $Q_{x,z}^J$ of γ_x^J and γ_z^J for $x, z \in \mathbb{S}$ that differ only at the site $j \in I$. We distinguish the following cases:

1. $d(j) = (l-1)q$ and $v(j) \neq v$;
2. $d(j) = (l-1)q$ and $v(j) = v$;
3. $(l-1)q + 1 \le d(j) \le lq \wedge \ell(i)$ and $v(j) \neq v$;
4. $d(j) = lq \wedge \ell(i) + 1$ and $v(j) \neq v$;
5. $d(j) = lq \wedge \ell(i) + 1$ and $v(j) = v$.

It is easily seen that γ_x^J does not depend on x^j except in one of the above cases. Thus when j satisfies none of the above conditions, we can set $C_{aj} = 0$ for $a \in J$.

Case 1. In this case, we must have $j \in c(i)$ with $v(j) \neq v$. Note that

$$\gamma_x^J(A) \ge \varepsilon^2 \frac{\int \mathbf{1}_A(x^J)\, q^v(x^{i-}, x^i) \prod_{a \in I_+ : J \cap c(a) \neq \varnothing} p^{v(a)}(x^{c(a)}, x^a) \prod_{b \in J} g^v(x^b, Y^b)\, \psi^v(dx^b)}{\int q^v(x^{i-}, x^i) \prod_{a \in I_+ : J \cap c(a) \neq \varnothing} p^{v(a)}(x^{c(a)}, x^a) \prod_{b \in J} g^v(x^b, Y^b)\, \psi^v(dx^b)},$$

where we define $i_- \in c(i)$ to be the (unique) child of i such that $v(i_-) = v(i)$. As the right hand side does not depend on x^j, we can construct a coupling $Q^J_{x,z}$ using Lemma C.15 such that $C_{aj} \le 1 - \varepsilon^2$ for every $a \in J$ and $x, z \in \mathbb{S}$ such that $x^{I\setminus\{j\}} = z^{I\setminus\{j\}}$.

Case 2. In this case we have $j = i_-$. Let us write $J = \{i_1, \dots, i_u\}$ where $u = \operatorname{card} J$ and $d(i_k) = (l-1)q + k$ for $k = 1, \dots, u$. Thus $i_1 = i$, and we define $i_0 = i_-$. Let us also write $\tilde{J}_k = \{i_k, \dots, i_u\}$. Then we can define the transition kernels on $\mathbb{X}^v$

$$P_{k,x}(\omega, A) = \frac{\int \mathbf{1}_A(x^{i_k})\, p^v(\omega x^{c(i_k)\setminus i_{k-1}}, x^{i_k}) \prod_{\tilde{J}_k \cap c(a) \neq \varnothing} p^{v(a)}(x^{c(a)}, x^a) \prod_{b \in \tilde{J}_k} g^v(x^b, Y^b)\, \psi^v(dx^b)}{\int p^v(\omega x^{c(i_k)\setminus i_{k-1}}, x^{i_k}) \prod_{\tilde{J}_k \cap c(a) \neq \varnothing} p^{v(a)}(x^{c(a)}, x^a) \prod_{b \in \tilde{J}_k} g^v(x^b, Y^b)\, \psi^v(dx^b)}$$

for $k = 1, \dots, u$. By construction, $P_{k,x}(x^{i_{k-1}}, dx^{i_k}) = \gamma^J_x(dx^{i_k} | x^{i_1}, \dots, x^{i_{k-1}})$, so we are in the setting of Lemma C.16. Moreover, we can estimate

$$P_{k,x}(\omega, A) \ge \varepsilon^2 \delta^2 \frac{\int \mathbf{1}_A(x^{i_k}) \prod_{\tilde{J}_k \cap c(a) \neq \varnothing} p^{v(a)}(x^{c(a)}, x^a) \prod_{b \in \tilde{J}_k} g^v(x^b, Y^b)\, \psi^v(dx^b)}{\int \prod_{\tilde{J}_k \cap c(a) \neq \varnothing} p^{v(a)}(x^{c(a)}, x^a) \prod_{b \in \tilde{J}_k} g^v(x^b, Y^b)\, \psi^v(dx^b)}.$$

Thus whenever $x, z \in \mathbb{S}$ satisfy $x^{I\setminus\{j\}} = z^{I\setminus\{j\}}$, we can construct a coupling $Q^J_{x,z}$ using Lemma C.16 such that $C_{i_k j} \le (1 - \varepsilon^2\delta^2)^k$ for every $k = 1, \dots, u$.

Case 3. In this case we have $j \in \bigcup_{a \in I_+ : J \cap c(a) \neq \varnothing} c(a)$ or $J \cap c(j) \neq \varnothing$, with $v(j) \neq v$. Let us note for future reference that there are at most $q\Delta^2$ such sites j. Now note that

$$\gamma^J_x(A) \ge \varepsilon^{2(\Delta+1)} \times \frac{\int \mathbf{1}_A(x^J)\, p^v(x^{c(i)}, x^i) \prod_{a \in I_+ : J \cap c(a) \neq \varnothing} \beta^a(x^{c(a)}, x^a) \prod_{b \in J} g^v(x^b, Y^b)\, \psi^v(dx^b)}{\int p^v(x^{c(i)}, x^i) \prod_{a \in I_+ : J \cap c(a) \neq \varnothing} \beta^a(x^{c(a)}, x^a) \prod_{b \in J} g^v(x^b, Y^b)\, \psi^v(dx^b)},$$

where we have defined $\beta^a(x^{c(a)}, x^a) = q^{v(a)}(x^{a-}, x^a)$ whenever $j = a$ or $j \in c(a)$, and $\beta^a(x^{c(a)}, x^a) = p^{v(a)}(x^{c(a)}, x^a)$ otherwise. The right hand side of this expression does not depend on x^j as the terms $q^{v(a)}(x^{a-}, x^a)$ for $v(a) \neq v$ cancel in the numerator and denominator. Thus whenever $x, z \in \mathbb{S}$ satisfy $x^{I\setminus\{j\}} = z^{I\setminus\{j\}}$, we can construct a coupling $Q^J_{x,z}$ using Lemma C.15 such that $C_{aj} \le 1 - \varepsilon^{2(\Delta+1)}$ for every $a \in J$.

Case 4. In this case $J \cap c(j) \neq \varnothing$ with $v(j) \neq v$. Note that

$$\gamma^J_x(A) \ge \varepsilon^2 \frac{\int \mathbf{1}_A(x^J)\, p^v(x^{c(i)}, x^i) \prod_{a \in I_+ : J \cap c(a) \neq \varnothing} \beta^a(x^{c(a)}, x^a) \prod_{b \in J} g^v(x^b, Y^b)\, \psi^v(dx^b)}{\int p^v(x^{c(i)}, x^i) \prod_{a \in I_+ : J \cap c(a) \neq \varnothing} \beta^a(x^{c(a)}, x^a) \prod_{b \in J} g^v(x^b, Y^b)\, \psi^v(dx^b)},$$

where $\beta^a(x^{c(a)}, x^a) = q^{v(a)}(x^{a-}, x^a)$ when $j = a$, and $\beta^a(x^{c(a)}, x^a) = p^{v(a)}(x^{c(a)}, x^a)$ otherwise. The right hand side does not depend on x^j as the term $q^{v(j)}(x^{j-}, x^j)$ cancels in the numerator and denominator. Thus whenever $x, z \in \mathbb{S}$ satisfy $x^{I\setminus\{j\}} = z^{I\setminus\{j\}}$, we can construct a coupling $Q^J_{x,z}$ using Lemma C.15 such that $C_{aj} \le 1 - \varepsilon^2$ for every $a \in J$.

Case 5. In this case we have $j_- \in J$. Note that the existence of such j necessarily implies that $\ell(i) > lq$ by the definition of J. We can therefore write $J = \{i_1, \ldots, i_q\}$ where $d(i_k) = lq - k + 1$ for $k = 1, \ldots, q$, and we define $i_0 = j$. Let us also define the sets $\tilde{J}_k = \{i_k, \ldots, i_q\}$. Then we can define the transition kernels on $\mathbb{X}^v$

$$P_{k,x}(\omega, A) = \frac{\int \mathbf{1}_A(x^{i_k})\, p^v(x^{c(i_q)}, x^{i_q}) \prod_{a \in I_+ : \tilde{J}_k \cap c(a) \neq \varnothing} \beta^a_\omega(x^{c(a)}, x^a) \prod_{b \in \tilde{J}_k} g^v(x^b, Y^b)\, \psi^v(dx^b)}{\int p^v(x^{c(i_q)}, x^{i_q}) \prod_{a \in I_+ : \tilde{J}_k \cap c(a) \neq \varnothing} \beta^a_\omega(x^{c(a)}, x^a) \prod_{b \in \tilde{J}_k} g^v(x^b, Y^b)\, \psi^v(dx^b)}$$

for $k = 1, \ldots, q$, where $\beta^a_\omega(x^{c(a)}, x^a) = p^v(x^{c(a)}, \omega)$ if $a = i_{k-1}$ and $\beta^a_\omega(x^{c(a)}, x^a) = p^{v(a)}(x^{c(a)}, x^a)$ otherwise. By construction $P_{k,x}(x^{i_{k-1}}, dx^{i_k}) = \gamma^J_x(dx^{i_k} | x^{i_1}, \ldots, x^{i_{k-1}})$, so we are in the setting of Lemma C.16. Moreover, we can estimate

$$P_{k,x}(\omega, A) \geq \varepsilon^2 \delta^2 \times \frac{\int \mathbf{1}_A(x^{i_k})\, p^v(x^{c(i_q)}, x^{i_q}) \prod_{a \in I_+ : \tilde{J}_k \cap c(a) \neq \varnothing} \beta^a(x^{c(a)}, x^a) \prod_{b \in \tilde{J}_k} g^v(x^b, Y^b)\, \psi^v(dx^b)}{\int p^v(x^{c(i_q)}, x^{i_q}) \prod_{a \in I_+ : \tilde{J}_k \cap c(a) \neq \varnothing} \beta^a(x^{c(a)}, x^a) \prod_{b \in \tilde{J}_k} g^v(x^b, Y^b)\, \psi^v(dx^b)},$$

where $\beta^a(x^{c(a)}, x^a) = 1$ if $a = i_{k-1}$ and $\beta^a(x^{c(a)}, x^a) = p^{v(a)}(x^{c(a)}, x^a)$ otherwise. Thus whenever $x, z \in \mathbb{S}$ satisfy $x^{I \setminus \{j\}} = z^{I \setminus \{j\}}$, we can construct a coupling $Q^J_{x,z}$ using Lemma C.16 such that $C_{i_k j} \leq (1 - \varepsilon^2 \delta^2)^k$ for every $k = 1, \ldots, q$.

We have now constructed coupled updates $Q^J_{x,z}$ for every pair $x, z \in \mathbb{S}$ that differ only at one point. Collecting the above bounds on the matrix C, we can estimate

$$\sum_{j \in I} e^{\beta |d(a) - d(j)|} C_{aj} \leq 3q\Delta^2 e^{\beta q}(1 - \varepsilon^{2(\Delta+1)}) + e^\beta (1 - \varepsilon^2 \delta^2) + e^{\beta q}(1 - \varepsilon^2 \delta^2)^q =: c$$

whenever $a \in J$, where we have used the convexity of the function $\alpha^{x+1} + \alpha^{q-x}$.

Up to this point we have considered an arbitrary block $J = J^i_l \in \mathcal{J}$ with $1 \leq l \leq \lceil n/q \rceil$. However, in the remaining case $l = 0$ it is easily seen that $\gamma^J_x = \delta_{\sigma^J}$ does not depend on x, so we can evidently set $C_{aj} = 0$ for $a \in J$. Thus we have shown that

$$\|C\|_{\infty, \beta m} := \max_{i \in I} \sum_{j \in I} e^{\beta m(i,j)} C_{ij} \leq c,$$

where we define the pseudometric $m(i, j) = |d(i) - d(j)|$. On the other hand, in the present setting it is evident that $\gamma^J_x = \tilde{\gamma}^J_x$ whenever $J = J^i_l \in \mathcal{J}$ with $1 \leq l \leq \lceil n/q \rceil$. We can therefore choose couplings $\hat{Q}^J_x$ such that $b_i \leq \mathbf{1}_{d(i)=0}$ for all $i \in I$. Substituting into the comparison theorem and arguing as in the proof of Theorem C.14 yields the estimate

$$\|\tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_1 \delta_\sigma - \tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_1 \delta_{\tilde{\sigma}}\|_J \leq \frac{2}{1 - c} \operatorname{card} J\, e^{-\beta n}.$$

Thus the proof is complete. □

Proposition C.18 provides control of the block filter as a function of time but not as a function of the initial conditions. The dependence on the initial conditions can however be incorporated *a posteriori* as in the proof of Proposition A.15. This yields the following result, which forms the basis for the proof of Theorem C.20 below.

Corollary C.19 (Block filter stability)**.** *Suppose there exist* $0 < \varepsilon, \delta < 1$ *such that*

$$\varepsilon q^v(x^v, z^v) \le p^v(x, z^v) \le \varepsilon^{-1} q^v(x^v, z^v),$$
$$\delta \le q^v(x^v, z^v) \le \delta^{-1}$$

for every $v \in V$ *and* $x, z \in \mathbb{X}$*, where* $q^v : \mathbb{X}^v \times \mathbb{X}^v \to \mathbb{R}_+$ *is a transition density with respect to* ψ^v*. Suppose also that we can choose* $q \in \mathbb{N}$ *and* $\beta > 0$ *such that*

$$c := 3q\Delta^2 e^{\beta q}(1 - \varepsilon^{2(\Delta+1)}) + e^{\beta}(1 - \varepsilon^2\delta^2) + e^{\beta q}(1 - \varepsilon^2\delta^2)^q < 1.$$

Let μ *and* ν *be (possibly random) probability measures on* $\mathbb{X}$ *of the form*

$$\mu = \bigotimes_{K \in \mathcal{K}} \mu^K, \qquad \nu = \bigotimes_{K \in \mathcal{K}} \nu^K.$$

Then we have

$$\|\tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_{s+1}\mu - \tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_{s+1}\nu\|_J \le \frac{2}{1-c} \operatorname{card} J \, e^{-\beta(n-s)},$$

as well as

$$\mathbf{E}[\|\tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_{s+1}\mu - \tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_{s+1}\nu\|_J^2]^{1/2}$$
$$\le \frac{2}{1-c} \frac{1}{(\varepsilon\delta)^{2|\mathcal{K}|_\infty}} \operatorname{card} J \, (e^{-\beta}\Delta_{\mathcal{K}})^{n-s} \max_{K \in \mathcal{K}} \mathbf{E}[\|\mu^K - \nu^K\|^2]^{1/2},$$

for every $s < n$*,* $K \in \mathcal{K}$ *and* $J \subseteq K$*.*

Proof. The proof is a direct adaptation of the proof of Proposition A.15. □

The block filter stability result in Appendix A is the only place in the proof of the variance bound where the inadequacy of the classical comparison theorem plays a role. Having exploited the generalized comparison Theorem 6.4 to extend the stability results in Appendix A to the present setting, we would therefore expect that the remainder of the proof of the variance bound follows verbatim from Appendix A. Unfortunately, however, there is a complication: the result of Corollary C.19 is not as powerful as the corresponding result in Appendix A. Note that the first (uniform) bound in Corollary C.19 decays exponentially in time n, but the second (initial condition dependent) bound only decays in n if it happens to be the case that $e^{-\beta}\Delta_{\mathcal{K}} < 1$. As in Appendix A both the spatial and temporal interactions were assumed to be sufficiently weak, we could assume that the latter was always the case. In the present setting, however, it is possible that $e^{-\beta}\Delta_{\mathcal{K}} \ge 1$ no matter how weak are the spatial correlations.

To surmount this problem, we will use a slightly different error decomposition than was used in Appendix A to complete the proof of the variance bound. The present approach is inspired by [15]. The price we pay is that the variance bound scales in the number of samples as $N^{-\gamma}$ where γ may be less than the optimal (by the central limit theorem) rate $\frac{1}{2}$. It is likely that a more sophisticated method of proof would yield

the optimal $N^{\frac{1}{2}}$ rate in the variance bound. However, let us note that in order to put the block particle filter to good use we must optimize over the size of the blocks in $\mathcal{K}$, and optimizing the error bound in Theorem 6.13 yields at best a rate of order $N^{-\alpha}$ for some constant α depending on the constants β_1, β_2. As the proof of Theorem 6.13 is not expected to yield realistic values for the constants β_1, β_2, the suboptimality of the variance rate γ does not significantly alter the practical conclusions that can be drawn from Theorem 6.13.

We now proceed to the variance bound. The following is the main result of this section.

Theorem C.20 (Variance term). *Suppose there exist* $0 < \varepsilon, \delta, \kappa < 1$ *such that*

$$\begin{aligned} \varepsilon q^v(x^v, z^v) \le p^v(x, z^v) \le \varepsilon^{-1} q^v(x^v, z^v), \\ \delta \le q^v(x^v, z^v) \le \delta^{-1}, \\ \kappa \le g^v(x^v, y^v) \le \kappa^{-1} \end{aligned}$$

for every $v \in V$, $x, z \in \mathbb{X}$, *and* $y \in \mathbb{Y}$, *where* $q^v : \mathbb{X}^v \times \mathbb{X}^v \to \mathbb{R}_+$ *is a transition density with respect to* ψ^v. *Suppose also that we can choose* $q \in \mathbb{N}$ *and* $\beta > 0$ *such that*

$$c := 3q\Delta^2 e^{\beta q}(1 - \varepsilon^{2(\Delta+1)}) + e^{\beta}(1 - \varepsilon^2\delta^2) + e^{\beta q}(1 - \varepsilon^2\delta^2)^q < 1.$$

Then for every $n \ge 0$, $\sigma \in \mathbb{X}$, $K \in \mathcal{K}$ *and* $J \subseteq K$, *the following hold:*

1. *If* $e^{-\beta}\Delta_{\mathcal{K}} < 1$, *we have*

$$\|\!|\tilde{\pi}_n^\sigma - \hat{\pi}_n^\sigma|\!\|_J \le \operatorname{card} J \, \frac{32\Delta_{\mathcal{K}}}{1-c} \, \frac{2 - e^{-\beta}\Delta_{\mathcal{K}}}{1 - e^{-\beta}\Delta_{\mathcal{K}}} \, \frac{(\varepsilon\delta\kappa^{\Delta_{\mathcal{K}}})^{-4|\mathcal{K}|_\infty}}{N^{\frac{1}{2}}}.$$

2. *If* $e^{-\beta}\Delta_{\mathcal{K}} = 1$, *we have*

$$\|\!|\tilde{\pi}_n^\sigma - \hat{\pi}_n^\sigma|\!\|_J \le \operatorname{card} J \, \frac{16\beta^{-1}\Delta_{\mathcal{K}}}{1-c} \, (\varepsilon\delta\kappa^{\Delta_{\mathcal{K}}})^{-4|\mathcal{K}|_\infty} \, \frac{3 + \log N}{N^{\frac{1}{2}}}.$$

3. *If* $e^{-\beta}\Delta_{\mathcal{K}} > 1$, *we have*

$$\|\!|\tilde{\pi}_n^\sigma - \hat{\pi}_n^\sigma|\!\|_J \le \operatorname{card} J \, \frac{32\Delta_{\mathcal{K}}}{1-c} \left\{ \frac{1}{e^{-\beta}\Delta_{\mathcal{K}} - 1} + 2 \right\} \frac{(\varepsilon\delta\kappa^{\Delta_{\mathcal{K}}})^{-4|\mathcal{K}|_\infty}}{N^{\frac{\beta}{2\log\Delta_{\mathcal{K}}}}}.$$

The proof of Theorem C.20 combines the stability bounds of Corollary C.19 and one-step bounds on the sampling error, Lemma A.17 and Proposition A.20, that can be used verbatim in the present setting. We recall the latter here for the reader's convenience.

Proposition C.21 (Sampling error). *Suppose there exist* $0 < \varepsilon, \delta, \kappa < 1$ *such that*

$$\begin{aligned} \varepsilon q^v(x^v, z^v) \le p^v(x, z^v) \le \varepsilon^{-1} q^v(x^v, z^v), \\ \delta \le q^v(x^v, z^v) \le \delta^{-1}, \\ \kappa \le g^v(x^v, y^v) \le \kappa^{-1} \end{aligned}$$

for every $v \in V$, $x, z \in \mathbb{X}$, and $y \in \mathbb{Y}$. Then we have

$$\max_{K\in\mathcal{K}} \|\!|\!| \tilde{\mathsf{F}}_n \hat{\pi}^\sigma_{n-1} - \hat{\mathsf{F}}_n \hat{\pi}^\sigma_{n-1} |\!|\!|_K \le \frac{2\kappa^{-2|\mathcal{K}|_\infty}}{N^{\frac12}}$$

and

$$\max_{K\in\mathcal{K}} \mathbf{E}[\|\tilde{\mathsf{F}}_{s+1}\tilde{\mathsf{F}}_s \hat{\pi}^\sigma_{s-1} - \tilde{\mathsf{F}}_{s+1}\hat{\mathsf{F}}_s \hat{\pi}^\sigma_{s-1}\|_K^2]^{1/2} \le \frac{16\Delta_{\mathcal{K}}(\varepsilon\delta)^{-2|\mathcal{K}|_\infty}\kappa^{-4|\mathcal{K}|_\infty\Delta_{\mathcal{K}}}}{N^{\frac12}}$$

for every $0 < s < n$ and $\sigma \in \mathbb{X}$.

Proof. Immediate from Lemma A.17 and Proposition A.20 upon replacing ε by $\varepsilon\delta$. □

We can now prove Theorem C.20.

Proof of Theorem C.20. We fix for the time being an integer $t \ge 1$ (we will optimize over t at the end of the proof). We argue differently when $n \le t$ and when $n > t$.

Suppose first that $n \le t$. In this case, we estimate

$$\begin{aligned}
|\!|\!| \tilde{\pi}^\sigma_n - \hat{\pi}^\sigma_n |\!|\!|_J &= |\!|\!| \tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_1 \delta_\sigma - \hat{\mathsf{F}}_n \cdots \hat{\mathsf{F}}_1 \delta_\sigma |\!|\!|_J \\
&\le \sum_{k=1}^{n} |\!|\!| \tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_{k+1}\tilde{\mathsf{F}}_k \hat{\pi}^\sigma_{k-1} - \tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_{k+1}\hat{\mathsf{F}}_k \hat{\pi}^\sigma_{k-1} |\!|\!|_J
\end{aligned}$$

using a telescoping sum and the triangle inequality. The term $k = n$ in the sum is estimated by the first bound in Proposition C.21, while the remaining terms are estimated by the second bound of Corollary C.19 and Proposition C.21, respectively. This yields

$$|\!|\!| \tilde{\pi}^\sigma_n - \hat{\pi}^\sigma_n |\!|\!|_J \le \operatorname{card} J \, \frac{32\Delta_{\mathcal{K}}}{1-c} \, \frac{(\varepsilon\delta\kappa^{\Delta_{\mathcal{K}}})^{-4|\mathcal{K}|_\infty}}{N^{\frac12}} \left\{ \frac{(e^{-\beta}\Delta_{\mathcal{K}})^{n-1} - 1}{e^{-\beta}\Delta_{\mathcal{K}} - 1} + 1 \right\}$$

(in the case $e^{-\beta}\Delta_{\mathcal{K}} = 1$, the quantity between the brackets $\{\cdot\}$ equals n).

Now suppose that $n > t$. Then we decompose the error as

$$\begin{aligned}
|\!|\!| \tilde{\pi}^\sigma_n - \hat{\pi}^\sigma_n |\!|\!|_J \le{}& |\!|\!| \tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_{n-t+1} \tilde{\pi}^\sigma_{n-t} - \tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_{n-t+1} \hat{\pi}^\sigma_{n-t} |\!|\!|_J \\
&+ \sum_{k=n-t+1}^{n} |\!|\!| \tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_{k+1}\tilde{\mathsf{F}}_k \hat{\pi}^\sigma_{k-1} - \tilde{\mathsf{F}}_n \cdots \tilde{\mathsf{F}}_{k+1}\hat{\mathsf{F}}_k \hat{\pi}^\sigma_{k-1} |\!|\!|_J,
\end{aligned}$$

that is, we develop the telescoping sum for t steps only. The first term is estimated by the first bound in Corollary C.19, while the sum is estimated as in the case $n \le t$. This yields

$$|\!|\!| \tilde{\pi}^\sigma_n - \hat{\pi}^\sigma_n |\!|\!|_J \le \frac{\operatorname{card} J}{1-c} \left[2e^{-\beta t} + \frac{32\Delta_{\mathcal{K}}(\varepsilon\delta\kappa^{\Delta_{\mathcal{K}}})^{-4|\mathcal{K}|_\infty}}{N^{\frac12}} \left\{ \frac{(e^{-\beta}\Delta_{\mathcal{K}})^{t-1} - 1}{e^{-\beta}\Delta_{\mathcal{K}} - 1} + 1 \right\} \right]$$

(in the case $e^{-\beta}\Delta_{\mathcal{K}} = 1$, the quantity between the brackets $\{\cdot\}$ equals t).

We now consider separately the three cases in the statement of the Theorem.

Case 1. In this case we choose $t = n$, and note that

$$\frac{(e^{-\beta}\Delta_{\mathcal{K}})^{n-1} - 1}{e^{-\beta}\Delta_{\mathcal{K}} - 1} + 1 \leq \frac{2 - e^{-\beta}\Delta_{\mathcal{K}}}{1 - e^{-\beta}\Delta_{\mathcal{K}}} \quad \text{for all } n \geq 1.$$

Thus the result follows from the first bound above.

Case 2. In this case we have

$$|||\tilde{\pi}_n^\sigma - \hat{\pi}_n^\sigma|||_J \leq \frac{\operatorname{card} J}{1-c}\left[2e^{-\beta t} + \frac{32\Delta_{\mathcal{K}}(\varepsilon\delta\kappa^{\Delta_{\mathcal{K}}})^{-4|\mathcal{K}|_\infty}}{N^{\frac{1}{2}}}\, t\right]$$

for all $t, n \geq 1$. Now choose $t = \lceil (2\beta)^{-1} \log N \rceil$. Then

$$|||\tilde{\pi}_n^\sigma - \hat{\pi}_n^\sigma|||_J \leq \frac{\operatorname{card} J}{1-c}\left[16\beta^{-1}\Delta_{\mathcal{K}}(\varepsilon\delta\kappa^{\Delta_{\mathcal{K}}})^{-4|\mathcal{K}|_\infty}\frac{\log N}{N^{\frac{1}{2}}} + \frac{34\Delta_{\mathcal{K}}(\varepsilon\delta\kappa^{\Delta_{\mathcal{K}}})^{-4|\mathcal{K}|_\infty}}{N^{\frac{1}{2}}}\right],$$

which readily yields the desired bound.

Case 3. In this case we have

$$|||\tilde{\pi}_n^\sigma - \hat{\pi}_n^\sigma|||_J \leq \frac{\operatorname{card} J}{1-c}\left[2e^{-\beta t} + \frac{32\Delta_{\mathcal{K}}(\varepsilon\delta\kappa^{\Delta_{\mathcal{K}}})^{-4|\mathcal{K}|_\infty}}{N^{\frac{1}{2}}}\left\{\frac{(e^{-\beta}\Delta_{\mathcal{K}})^{t-1} - 1}{e^{-\beta}\Delta_{\mathcal{K}} - 1} + 1\right\}\right]$$

for all $t, n \geq 1$. Now choose $t = \left\lceil \frac{\log N}{2\log \Delta_{\mathcal{K}}} \right\rceil$. Then

$$|||\tilde{\pi}_n^\sigma - \hat{\pi}_n^\sigma|||_J \leq \operatorname{card} J\, \frac{32\Delta_{\mathcal{K}}}{1-c}\left\{\frac{1}{e^{-\beta}\Delta_{\mathcal{K}} - 1} + 2\right\}\frac{(\varepsilon\delta\kappa^{\Delta_{\mathcal{K}}})^{-4|\mathcal{K}|_\infty}}{N^{\frac{\beta}{2\log\Delta_{\mathcal{K}}}}},$$

and the proof is complete. □

The conclusion of Theorem 6.13 now follows readily from Theorems C.14 and C.20. We must only check that the assumptions Theorems C.14 and C.20 are satisfied. The assumption of Theorem C.14 is slightly stronger than that of Theorem C.20, so it suffices to consider the former. To this end, fix $0 < \delta < 1$, and choose $q \in \mathbb{N}$ such that

$$1 - \delta^2 + (1-\delta^2)^q < 1.$$

Then we may evidently choose $0 < \varepsilon_0 < 1$, depending on δ and Δ only, such that

$$3q\Delta^2(1 - \varepsilon^{2(\Delta+1)}) + 1 - \varepsilon^2\delta^2 + (1-\varepsilon^2\delta^2)^q < 1$$

for all $\varepsilon_0 < \varepsilon \leq 1$. This is the constant ε_0 that appears in the statement of Theorem 6.13. Finally, it is now clear that we can choose $\beta > 0$ sufficiently close to zero (depending on $\delta, \varepsilon, r, \Delta$ only) such that $c < 1$. Thus the proof of Theorem 6.13 is complete.

Appendix D

Nonlinear filtering in infinite dimension: proofs

This appendix is devoted to the proof of Theorem 7.7. The proof relies on standard tools from statistical mechanics [7, 27]: a Peierls argument for the low noise regime and a Dobrushin contraction method for the high noise regime.

D.1 Proof of Theorem 7.7: 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 $\mathbf{E}(X_k^0|Y_1,\ldots,Y_k) = \mathbf{E}(-X_k^0|Y_1,\ldots,Y_k)$, and we therefore have

$$\mathbf{E}(X_k^0|Y_1,\ldots,Y_k) = 0 \quad \text{for all } k \geq 1.$$

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

$$\inf_{k\geq 1} \mathbf{E}\,|\mathbf{E}(X_k^0|X_0,Y_1,\ldots,Y_k)| > 0.$$

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

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

$$\mathbf{E}(X_k^0|X_0,Y_1,\ldots,Y_k,\{X_1^v,\ldots,X_k^v : |v| > m\}) \xrightarrow{m\to\infty} \mathbf{E}(X_k^0|X_0,Y_1,\ldots,Y_k) \quad a.s.$$

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

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

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 $\hat{Y}_\ell^m$ and $\hat{Y}_\ell^{-m-1}$, $1 \le \ell \le k$ 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

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

We therefore obtain using the Bayes formula (Theorem 2.7)

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

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

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

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

for every $m \ge 0$. Thus $(Z^m)_{m \le 0}$ is a Markov chain under any regular version of the conditional distribution $\mathbf{P}(\,\cdot\,|X_0, Y_1, \ldots, Y_k)$ (almost surely with respect to the realization of $X_0, Y_1, \ldots, Y_k$). Moreover, the above estimate shows that the (random) transition kernels of this Markov chain satisfy the Doeblin condition [38, Theorem 16.2.4], so

$$\begin{aligned}
|\mathbf{E}(X_k^0|X_0, Y_1, \ldots, Y_k, \{X_1^v, \ldots, X_k^v : |v| > m\}) - \mathbf{E}(X_k^0|X_0, Y_1, \ldots, Y_k)| \\
\le 2(1 - e^{-4\beta k})^{m+1}
\end{aligned}$$

for all $m \ge 0$. This completes the proof. □

Lemma D.1 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 7.1), so we will assume without loss of generality in the sequel that $0 < p \le 1/2$. Applying Lemma D.1, it follows that in order to prove that the filter is not stable, it suffices to show that

$$\inf_{k,m \ge 1} \mathbf{E}\,|\mathbf{E}(X_k^0|X_0, Y_1, \ldots, Y_k, \{X_1^v, \ldots, X_k^v : |v| > m\})| > 0.$$

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 $0 \le \ell \le k$ and $|v| \le m + 1$. 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 D.2. *Suppose that $0 < p \le 1/2$, and let $k, m \ge 1$. Define the quantities $\beta := \log\sqrt{(1-p)/p}$, $J := [1, k] \times [-m, m]$, 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

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

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

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

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

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

where $|A|$ denotes the cardinality of a set A. The result now follows readily from the Bayes formula (Theorem 2.7) and the fact that $Y^{qr} = X^q X^r \xi^{qr}$ by construction. □

Lemma D.2 shows that the distribution $\mathbf{P}(\cdot\,|X_0, Y_1, \ldots, Y_k, \{X_1^v, \ldots, X_k^v : |v| > m\})$ 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 [7, section 6.4]. For concreteness, we include the requisite arguments in the present setting, which completes the proof.

Proposition D.3. *There exists an absolute constant $0 < p_\star < 1/2$ such that*

$$\mathbf{E}\,|\mathbf{E}(X_k^0 | X_0, Y_1, \ldots, Y_k, \{X_1^v, \ldots, X_k^v : |v| > m\})| \ge \tfrac{1}{4}$$

for every $k, m \ge 1$ 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 $\mathbf{P}(X_k^0 \in \cdot\,|X_0, Y_1, \ldots, Y_k, \{X_1^v, \ldots, X_k^v : |v| > m\})$ assigns a large probability to the actually realized value of X_k^0 at least half of the time (recall Lemma D.2). 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

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

where we have used Lemma D.2 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 above claim. To this end, we use a Peierls argument. Fix for the time being a configuration $z \in \{-1, 1\}^J$. For any $J' \subseteq J$, define the boundary edges

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

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 $r \in J \backslash J'$. 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 $r \in J \backslash J'$. 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 D.2. Let us make this precise.

Lemma D.4. *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

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

Then we evidently have by Lemma D.2

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

An elementary computation shows that

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

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. □

Lemma D.4 allows us to estimate

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

Using a standard combinatorial result [27, 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

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

we can conclude that

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

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

D.2 Proof of Theorem 7.7: 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 D.5. *Suppose that*

$$\mathbf{E}\,|\mathbf{E}(f(X_k^{-m},\ldots,X_k^m)|X_0,Y_1,\ldots,Y_k)-\mathbf{E}(f(X_k^{-m},\ldots,X_k^m)|Y_1,\ldots,Y_k)| \xrightarrow{k\to\infty} 0$$

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

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

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

We can estimate

$$\begin{aligned}
&\mathbf{E}\,|\mathbf{P}(X_k\in A|X_0,Y_1,\ldots,Y_k)-\mathbf{P}(X_k\in A|Y_1,\ldots,Y_k)|\\
&\quad\le 2\,\mathbf{E}\,|f_m(X_k^{-m},\ldots,X_k^m)-\mathbf{1}_A(X_k)|\\
&\qquad+\mathbf{E}\,|\mathbf{E}(f_m(X_k^{-m},\ldots,X_k^m)|X_0,Y_1,\ldots,Y_k)-\mathbf{E}(f_m(X_k^{-m},\ldots,X_k^m)|Y_1,\ldots,Y_k)|.
\end{aligned}$$

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

$$\limsup_{k\to\infty}\mathbf{E}\,|\mathbf{P}(X_k\in A|X_0,Y_1,\ldots,Y_k)-\mathbf{P}(X_k\in A|Y_1,\ldots,Y_k)|\le 2\,\mathbf{E}\,|F_m-\mathbf{1}_A(X_0)|.$$

Letting $m\to\infty$ and using the martingale convergence theorem concludes the proof. □

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 (Theorem 2.11), which we state here in a convenient form.

Theorem D.6 (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*

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

Define

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

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

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

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

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 D.7. *Fix any version of the regular conditional probabilities*

$$\mu_{X,Y} := \mathbf{P}(X_0, \ldots, X_k \in \cdot\,|X_0, Y_1, \ldots, Y_k), \qquad \nu_Y := \mathbf{P}(X_0, \ldots, X_k \in \cdot\,|Y_1, \ldots, Y_k).$$

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 $v \in \mathbb{Z}$,

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

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 (we use the notation $Y_{1:k} = Y_1, \ldots, Y_k$)

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

see [63, p. 95–96] or [52, Lemma 3.4]. That each statement in the Lemma holds for $\mathbf{P}$-a.e. (x,y) can therefore be read off from Lemma D.2. As there are countably many statements, they can be assumed to hold simultaneously on a set A of unit measure. □

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

Proposition D.8. *There exists an absolute constant $0 < p^\star < 1/2$ such that*

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

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

Proof. We apply Theorem D.6 with $I = \{0, \ldots, k\} \times \mathbb{Z}$ and $\mu = \mu_{x,y}$, $\nu = \nu_y$ as defined in Lemma D.7. 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

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

by Theorem D.6 provided that the condition on the matrix C is satisfied.

We proceed to estimate the matrix C using Lemma D.7. Evidently

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

On the other hand, note that by Lemma D.7

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

so we can estimate

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

It follows readily that

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

We can now evidently choose $0 < p^\star < 1/2$ such that $4e \tanh(4\beta) < 1/2$ for $p^\star < p \le 1/2$. Then the condition of Theorem D.6 is satisfied. Moreover, as $\|\cdot\|_*$ is a matrix norm

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

Thus we obtain

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

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

Bibliography

[1] A. Apte, C. K. R. T. Jones, A. M. Stuart, and J. Voss. Data assimilation: mathematical and statistical perspectives. *Int. J. Numer. Meth. Fluids*, 56:1033–1046, 2008.

[2] Søren Asmussen and Peter W. Glynn. *Stochastic Simulation: Algorithms and Analysis: Algorithms and Analysis*. Stochastic Modelling and Applied Probability. Springer, 2007.

[3] Alexandros Beskos, Dan Crisan, Ajay Jasra, and Nick Whiteley. Error bounds and normalizing constants for sequential Monte Carlo in high dimensions, 2012. Preprint arxiv:1112.1544v1.

[4] Peter Bickel, Bo Li, and Thomas Bengtsson. Sharp failure rates for the bootstrap particle filter in high dimensions. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, volume 3 of *Inst. Math. Stat. Collect.*, pages 318–329. Inst. Math. Statist., Beachwood, OH, 2008.

[5] David Blackwell. The entropy of functions of finite-state Markov chains. In *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*, pages 13–20. Publishing House of the Czechoslovak Academy of Sciences, Prague, 1957.

[6] Michael Blank. *Discreteness and continuity in problems of chaotic dynamics*, volume 161 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1997. Translated from the Russian manuscript by the author.

[7] Anton Bovier. *Statistical mechanics of disordered systems*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2006. A mathematical perspective.

[8] Olivier Cappé, Eric Moulines, and Tobias Rydén. *Inference in hidden Markov models*. Springer Series in Statistics. Springer, New York, 2005.

[9] Pavel Chigansky and Ramon van Handel. A complete solution to Blackwell's unique ergodicity problem for hidden Markov chains. *Ann. Appl. Probab.*, 20(6):2318–2345, 2010.

[10] Alexandre J. Chorin and Matthias Morzfeld. Conditions for successful data assimilation, 2013. Preprint arXiv:1303.2714.

[11] Erhan Çınlar. *Probability and Stochastics*. Graduate Texts in Mathematics. Springer, 2011.

[12] J. P. Conze. Entropie d'un groupe abélien de transformations. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 25:11–30, 1972/73.

[13] Dan Crisan and Boris Rozovskiĭ, editors. *The Oxford handbook of nonlinear filtering*. Oxford University Press, Oxford, 2011.

[14] T. de la Rue, R. Fernández, and A. D. Sokal. How to clean a dirty floor: probabilistic potential theory and the Dobrushin uniqueness theorem. *Markov Process. Related Fields*, 14(1):1–78, 2008.

[15] Pierre Del Moral and Alice Guionnet. On the stability of interacting processes with applications to filtering and genetic algorithms. *Ann. Inst. H. Poincaré Probab. Statist.*, 37(2):155–194, 2001.

[16] R. L. Dobrushin, V. I. Kryukov, and A. L. Toom, editors. *Stochastic Cellular Systems: Ergodicity, Memory, Morphogenesis*. Manchester University Press, Manchester, 1990.

[17] Roland L. Dobrushin and Senya B. Shlosman. Constructive criterion for the uniqueness of Gibbs field. In *Statistical physics and dynamical systems (Köszeg, 1984)*, volume 10 of *Progr. Phys.*, pages 347–370. Birkhäuser Boston, Boston, MA, 1985.

[18] R. L. Dobrušin. Definition of a system of random variables by means of conditional distributions. *Teor. Verojatnost. i Primenen.*, 15:469–497, 1970.

[19] Arnaud Doucet and Adam M. Johansen. A tutorial on particle filtering and smoothing: fifteen years later, 2011.

[20] R. M. Dudley. *Real analysis and probability*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2002. Revised reprint of the 1989 original.

[21] Martin Dyer, Leslie Ann Goldberg, and Mark Jerrum. Dobrushin conditions and systematic scan. *Combin. Probab. Comput.*, 17(6):761–779, 2008.

[22] Martin Dyer, Leslie Ann Goldberg, and Mark Jerrum. Matrix norms and rapid mixing for spin systems. *Ann. Appl. Probab.*, 19(1):71–107, 2009.

[23] Hans Föllmer. Tail structure of Markov chains on infinite product spaces. *Z. Wahrsch. Verw. Gebiete*, 50(3):273–285, 1979.

[24] Hans Föllmer. A covariance estimate for Gibbs measures. *J. Funct. Anal.*, 46(3):387–395, 1982.

[25] Hans Föllmer. Random fields and diffusion processes. In *École d'Été de Probabilités de Saint-Flour XV–XVII, 1985–87*, volume 1362 of *Lecture Notes in Math.*, pages 101–203. Springer, Berlin, 1988.

[26] Arnoldo Frigessi, Fabio Martinelli, and Julian Stander. Computational complexity of Markov chain Monte Carlo methods for finite Markov random fields. *Biometrika*, 84(1):1–18, 1997.

[27] Hans-Otto Georgii. *Gibbs measures and phase transitions*, volume 9 of *de Gruyter Studies in Mathematics*. Walter de Gruyter & Co., Berlin, second edition, 2011.

[28] N.J. Gordon, D.J. Salmond, and A.F.M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEEE Proceedings F, Radar and Signal Processing*, 140(2):107–113, 1993.

[29] A. Guionnet and B. Zegarlinski. Lectures on logarithmic Sobolev inequalities. In *Séminaire de Probabilités, XXXVI*, volume 1801 of *Lecture Notes in Math.*, pages 1–134. Springer, Berlin, 2003.

[30] J. E. Handschin and D. Q. Mayne. Monte Carlo techniques to estimate the conditional expectation in multi-stage non-linear filtering. *International Journal of Control*, 9(5):547–559, 1969.

[31] Olav Kallenberg. *Foundations of modern probability*. Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2002.

[32] Christof Külske. Concentration inequalities for functions of Gibbs fields with application to diffraction and random Gibbs measures. *Comm. Math. Phys.*, 239(1-2):29–51, 2003.

[33] Hiroshi Kunita. Asymptotic behavior of the nonlinear filtering errors of Markov processes. *J. Multivariate Anal.*, 1:365–393, 1971.

[34] K. J. H. Law and A. M. Stuart. Evaluating data assimilation algorithms. *Mon. Weather Rev.*, 140:3757–3782, 2012.

[35] Joel L. Lebowitz, Christian Maes, and Eugene R. Speer. Statistical mechanics of probabilistic cellular automata. *J. Statist. Phys.*, 59(1-2):117–170, 1990.

[36] Thomas M. Liggett. *Interacting particle systems*. Classics in Mathematics. Springer-Verlag, Berlin, 2005. Reprint of the 1985 original.

[37] Andrew J. Majda and John Harlim. *Filtering complex turbulent systems*. Cambridge University Press, Cambridge, 2012.

[38] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Communications and Control Engineering Series. Springer-Verlag London Ltd., London, 1993.

[39] Elchanan Mossel and Allan Sly. Exact thresholds for Ising–Gibbs samplers on general graphs. *The Annals of Probability*, 41(1):294–328, 01 2013.

[40] Patrick Rebeschini and Ramon van Handel. Can local particle filters beat the curse of dimensionality? 2013. Preprint arXiv:1301.6585.

[41] Patrick Rebeschini and Ramon van Handel. Comparison theorems for Gibbs measures. 2013. Preprint arXiv:1308.4117.

[42] Patrick Rebeschini and Ramon van Handel. Phase transitions in nonlinear filtering. 2014. Preprint arXiv:1401.6450.

[43] H. L. Royden. *Real analysis*. Macmillan Publishing Company, New York, third edition, 1988.

[44] A.N. Shiryaev. *Probability*. Graduate Texts in Mathematics. Springer, 1996.

[45] Barry Simon. *The statistical mechanics of lattice gases. Vol. I*. Princeton Series in Physics. Princeton University Press, Princeton, NJ, 1993.

[46] Chris Snyder. Particle filters, the "optimal" proposal and high-dimensional systems. Proceedings, ECMWF Seminar on Data Assimilation for Atmosphere and Ocean, 6-9 September 2011.

[47] Chris Snyder, Thomas Bengtsson, Peter Bickel, and Jeff Anderson. Obstacles to high-dimensional particle filtering. *Mon. Weather Rev.*, 136:4629–4640, 2008.

[48] Łukasz Stettner. On invariant measures of filtering processes. In *Stochastic differential systems (Bad Honnef, 1988)*, volume 126 of *Lecture Notes in Control and Inform. Sci.*, pages 279–292. Springer, Berlin, 1989.

[49] Andrew M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numer.*, 19:451–559, 2010.

[50] Sekhar C. Tatikonda and Michael I. Jordan. Loopy belief propagation and Gibbs measures. In *Proc. UAI*, volume 18, pages 493–500, 2002.

[51] Xin Thomson Tong and Ramon van Handel. Ergodicity and stability of the conditional distributions of nondegenerate Markov chains. *Ann. Appl. Probab.*, 22(4):1495–1540, 2012.

[52] Xin Thomson Tong and Ramon van Handel. Conditional ergodicity in infinite dimension. *Annals of Probability, to appear*, 2013.

[53] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[54] Ramon van Handel. Discrete time nonlinear filters with informative observations are stable. *Electron. Commun. Probab.*, 13:562–575, 2008.

[55] Ramon van Handel. Hidden Markov models. Unpublished notes, 2008.

[56] Ramon van Handel. Observability and nonlinear filtering. *Probab. Theory Related Fields*, 145(1-2):35–74, 2009.

[57] Ramon van Handel. The stability of conditional Markov processes and Markov chains in random environments. *Ann. Probab.*, 37(5):1876–1925, 2009.

[58] Ramon van Handel. Uniform observability of hidden Markov models and filter stability for unstable signals. *Ann. Appl. Probab.*, 19(3):1172–1199, 2009.

[59] Ramon van Handel. On the exchange of intersection and supremum of σ-fields in filtering theory. *Israel J. Math.*, 192(2):763–784, 2012.

[60] Peter Jan van Leeuwen. Particle filtering in geophysical systems. *Mon. Weather Rev.*, 137:4089–4114, 2009.

[61] Eric Vanden-Eijnden and Jonathan Weare. Data assimilation in the low noise regime with application to the kuroshio. 2014. Preprint arXiv:1202.4952.

[62] Cédric Villani. *Optimal transport*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. Old and new.

[63] Heinrich von Weizsäcker. Exchanging the order of taking suprema and countable intersections of σ-algebras. *Ann. Inst. H. Poincaré Sect. B (N.S.)*, 19(1):91–100, 1983.

[64] Dror Weitz. Combinatorial criteria for uniqueness of Gibbs measures. *Random Structures & Algorithms*, 27(4):445–475, 2005.

[65] David Williams. *Probability with Martingales*. Cambridge mathematical textbooks. Cambridge University Press, 1991.

[66] Gerhard Winkler. *Image analysis, random fields and Markov chain Monte Carlo methods*, volume 27 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, second edition, 2003.

[67] Liming Wu. Poincaré and transportation inequalities for Gibbs measures under the Dobrushin uniqueness condition. *Ann. Probab.*, 34(5):1960–1989, 2006.