Whiteley, N., Jones, M. W., & Domanski, A. P. F. (2024). The infinite Viterbi alignment and decay-convexity. *Bernoulli*, *30*(1), 252-277. https://doi.org/10.48550/arXiv.1810.04115, https://doi.org/10.3150/23-BEJ1596

License (if available):
CC BY

Link to published version (if available):
10.48550/arXiv.1810.04115
10.3150/23-BEJ1596

Link to publication record in Explore Bristol Research
PDF-document

This is the accepted author manuscript (AAM) of the article which has been made Open Access under the University of Bristol's Scholarly Works Policy. The final published version (Version of Record) can be found on the publisher's website. The copyright of any third-party content, such as images, remains with the copyright holder.

University of Bristol - Explore Bristol Research
General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/

arXiv:1810.04115v5 [math.PR] 13 Feb 2023

The infinite Viterbi alignment and decay-convexity

Nick Whiteley*, Matt W. Jones †, Aleks Domanski‡

February 14, 2023

Abstract

The infinite Viterbi alignment is the limiting maximum a-posteriori estimate of the unobserved path in a hidden Markov model as the length of the time horizon grows. For models on state-space $\mathbb{R}^d$ satisfying a new "decay-convexity" condition, we develop an approach to existence of the infinite Viterbi alignment in an infinite dimensional Hilbert space. Quantitative bounds on the distance to the Viterbi process, which are the first of their kind, are derived and used to illustrate how approximate estimation via parallelization can be accurate and scaleable to high-dimensional problems because the rate of convergence to the infinite Viterbi alignment does not necessarily depend on d. The results are applied to approximate estimation via parallelization and a model of neural population activity.

1 Introduction

Let the *signal* $(X_n)_{n\geq 0}$ be a Markov chain with state space $\mathbb{R}^d$ whose initial distribution and transition kernel admit densities $\mu(x)$ and $f(x,x')$ with respect to d-dimensional Lebesgue measure. Let $(Y_n)_{n\geq 0}$, called the *observations*, be each valued in a measurable space $(\mathbb{Y},\mathcal{Y})$, conditionally independent given $(X_n)_{n\geq 0}$ and such that for any $A\in\mathcal{Y}$, the conditional probability of $\{Y_n\in A\}$ given $(X_n)_{n\geq 0}$ can be written in the form $\int_A g(X_n,y)\rho(\mathrm{d}y)$, where $g:\mathbb{R}^d\times\mathbb{Y}\to[0,+\infty)$ and ρ is a measure on $\mathcal{Y}$. Models of this form, under the names hidden Markov models (HMMs) or state space models, are applied in a huge variety of fields including econometrics, engineering, ecology, machine learning and neuroscience [20, 9, 8]. With:

$$U_n(x_0,\ldots,x_n,y_0,\ldots,y_n) \coloneqq -\log\mu(x_0)-\log g(x_0,y_0) - \sum_{m=1}^{n}\log f(x_{m-1},x_m) - \sum_{m=1}^{n}\log g(x_m,y_m), \tag{1}$$

the maximum a-posteriori path estimation problem given $(y_0,\ldots,y_n)$ is to find:

$$(\xi_{n,0},\ldots,\xi_{n,n}) \coloneqq \underset{x_0,\ldots,x_n}{\arg\min}\ U_n(x_0,\ldots,x_n,y_0,\ldots,y_n). \tag{2}$$

The dependence of $\xi_{n,0},\ldots,\xi_{n,n}$ on $y_0,\ldots,y_n$ is not shown in the notation. With $y\coloneqq(y_0,y_1,\ldots)\in\mathbb{Y}^{\mathbb{N}}$, the *infinite Viterbi alignment* is a sequence $\xi_\infty=(\xi_{\infty,n})_{n\geq 0}$ such that for any $m\in\mathbb{N}_0$,

$$(\xi_{\infty,0},\ldots,\xi_{\infty,m}) = \lim_{n\to\infty}(\xi_{n,0},\ldots,\xi_{n,m}), \quad \text{for} \quad \mathbf{P}_Y\text{-almost all } y, \tag{3}$$

where in general $\xi_\infty=(\xi_{\infty,n})_{n\geq 0}$ is an unknown function of the infinite observation sequence $y=(y_0,y_1,\ldots)$ and $\mathbf{P}_Y$ is some probability measure on $\mathcal{Y}^{\otimes\mathbb{N}}$. In the "well-specified" case, where $\mathbf{P}_Y$ is the probability law of the observation process $Y_0,Y_1,\ldots$ induced by the same μ, f and g defining each function U_n in (1), the infinite Viterbi alignment as a function of $Y_0,Y_1,\ldots$ is called the *Viterbi process*.

Whilst the existence and uniqueness of the left hand side of (2) can be addressed directly in terms of μ, f, and g, the existence and uniqueness of the infinite Viterbi alignment and Viterbi process are less obvious. Studies and applications to date have mostly focused on the case where the state space of the signal is a discrete finite set [2, 1, 12, 15, 14, 13] and there the convergence in (3) is with respect to

*School of Mathematics, University of Bristol and the Alan Turing Institute.
†School of Physiology, Pharmacology and Neuroscience, University of Bristol.
‡School of Physiology, Pharmacology and Neuroscience, University of Bristol and the Alan Turing Institute.

the discrete metric. In this discrete setting the infinite Viterbi alignment and Viterbi process have been also been studied recently for pairwise Markov models [16, 17]. However, the case of HMMs with signal state-space $\mathbb{R}^d$ is considerably less well-understood: the only work known to the authors which considers this case is [3], where convergence in (3) is with respect to Euclidean distance, and proofs are given only for the case $d = 1$. The work [3] considered two approaches to existence of the Viterbi process, one based on regeneration times which is reminiscent of ideas used in the discrete case, and one involving dynamic programming operators.

The assumptions in the present work are most directly comparable to those used in the dynamic programming approach of [3]—we shall discuss this in more detail in section 4.9—but our setup and proof techniques are quite different: we develop a new framework in which the infinite Viterbi alignment is an element of an infinite dimensional Hilbert space, $l_2(\gamma)$, where $\gamma \in (0, 1]$ is a parameter related to the rate of convergence to the infinite Viterbi alignment. This approach has several benefits. Firstly, it allows interpretable quantitative bounds to be obtained which measure the distance to the infinite Viterbi alignment in a norm on $l_2(\gamma)$ which gives a stronger notion of convergence than the pointwise convergence in (3). Secondly, via a new "decay-convexity" condition our approach provides a new characterization of the Viterbi process as the fixed point of an infinite dimensional ordinary differential equation which arises in the limit $n \to \infty$. Thirdly, our analysis conveniently allows us to handle misspecified hidden Markov models: the situation in which the probability measure $\mathbf{P}_Y$ in (3) need not be the probability law of $Y_0, Y_1, \ldots$ under the HMM specified by μ, f, g, nor necessarily of the law of the observation process under any HMM. Misspecification is a topic of interest in the study of asymptotic properties of maximum likelihood estimators [6] and forgetting of initial conditions [7][Remark 2] for HMMs. It is important in the present context because in practice there is usually some modelling error, and one computes maximum a-posteriori estimator as in (2) with little if any knowledge of the probability law of the observation process.

A shortcoming of the present work is that we do not investigate ergodic or regenerative properties of the Viterbi process, as conducted in the discrete case in, for example, [1]. We leave this as a topic for future research. Another restriction of our approach is that we consider only situations in which the infinite Viterbi alignment, when it exists, is unique, reflecting the fact that our conditions are related to convexity of $U_n(\cdot, y)$. The question of whether our approach can be extended to situations in which there is non-uniqueness is left for future research.

Existence of the infinite Viterbi alignment can be understood as meaning that maximum a-posteriori state estimates have vanishing dependence on the observations in the distant future. From a practical point of view, this is important because it indicates that such future observations may be largely irrelevant to estimation at the present time and so can be safely ignored, resulting in computational savings. This idea can be extended to design a scheme for parallelized approximate computation of $(\xi_{n,0}, \ldots, \xi_{n,n})$ to which we shall apply our quantitative convergence results.

We start with definitions and Condition 1 concerning μ, f, g in section 2, leading to the statements of Theorems 1 and 2, which establish quantitative bounds on the differences between maximum a-posteriori estimators as a function of the observation sequence $y = (y_0, y_1, \ldots)$. Section 4 maps out and demonstrates the two main steps involved in applying Theorem 1 or 2, and the use of these theorems to establish the existence of the infinite Viterbi alignment. This is achieved via some intermediate results for classes of HMMs, combined with example calculations for specific HMMs to illustrate the reasoning involved. Section 4 also contains discussion of how the quantitative $l_2(\gamma)$ bounds can be used to analyze the error associated with the aforementioned parallelized approximate estimation scheme, and an application with numerical results for a model of neural population activity.

2 Definitions

With $d \in \mathbb{N}$ fixed throughout, when x is a point in $\mathbb{R}^{\mathbb{N}}$, we associate with it the vectors $x_0, x_1, \ldots$, each in $\mathbb{R}^d$, such that $x = [x_0^{\mathrm{T}}\, x_1^{\mathrm{T}} \ldots]^{\mathrm{T}}$. With $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ the Euclidean inner product and norm on $\mathbb{R}^d$, define the inner product and norm on $\mathbb{R}^{\mathbb{N}}$ associated with a given $\gamma \in (0, 1]$,

$$\langle x, x' \rangle_\gamma := \sum_{n=0}^{\infty} \gamma^n \langle x_n, x'_n \rangle, \qquad \|x\|_\gamma = \langle x, x \rangle_\gamma^{1/2} := \left(\sum_{n=0}^{\infty} \gamma^n \|x_n\|^2 \right)^{1/2}.$$

Let $l_2(\gamma)$ be the Hilbert space consisting of the set $\{x \in \mathbb{R}^{\mathbb{N}} : \|x\|_\gamma < \infty\}$ equipped with the inner-product $\langle \cdot, \cdot \rangle_\gamma$ and the usual element-wise addition and scalar multiplication of vectors over field $\mathbb{R}$. For each

$n \geq 0$, $l_2^n(\gamma)$ denotes the subspace consisting of those $x \in l_2(\gamma)$ such that $x_m = 0$ for $m > n$, with the convention that $l_2^\infty(\gamma) \equiv l_2(\gamma)$. Note that for $n < \infty$ the set of vectors $l_2^n(\gamma)$ does not actually depend on γ; the notation $l_2^n(\gamma)$ is used as a reminder that $l_2^n(\gamma)$ is a subspace of $l_2(\gamma)$. Let $\|x\|_{\gamma,n} := \left(\sum_{m=0}^\infty \gamma^{|m-n|}\|x_m\|^2\right)^{1/2}$. When $y \in \mathbb{Y}^\mathbb{N}$ we shall identify $y = (y_0, y_1, \ldots)$ where each $y_n \in \mathbb{Y}$, stopping short of explicitly regarding y as a vector since for our main results we shall not need to assume that $\mathbb{Y}$ or $\mathbb{Y}^\mathbb{N}$ is a vector space.

For $x \in \mathbb{R}^\mathbb{N}$ and $y \in \mathbb{Y}^\mathbb{N}$ define

$$\phi_n(x,y) := \log f(x_{n-1}, x_n) + \log f(x_n, x_{n+1}) + \log g(x_n, y_n), \quad n \geq 1, \tag{4}$$

$$\tilde{\phi}_n(x,y) := \begin{cases} \log \mu(x_0) + \log f(x_0, x_1) + \log g(x_0, y_0), & n = 0, \\ \log f(x_{n-1}, x_n) + \log g(x_n, y_n), & n \geq 1. \end{cases} \tag{5}$$

and for each $n \geq 0$ let $\nabla_n \phi_n(x,y)$ and $\nabla_n \tilde{\phi}_n(x,y)$ be the vectors in $\mathbb{R}^d$ whose ith entries are the partial derivatives of $\phi_n(x,y)$ and $\tilde{\phi}_n(x,y)$ with respect to the ith entry of x_n (the existence of such derivatives is part of Condition 1 below).

For each $n \geq 0$, define the vector field $\nabla U_n : \mathbb{R}^\mathbb{N} \times \mathbb{Y}^\mathbb{N} \to \mathbb{R}^\mathbb{N}$,

$$\nabla U_n(x,y) := \\ -\left[\nabla_0 \tilde{\phi}_0(x,y)^\mathrm{T},\ \nabla_1 \phi_1(x,y)^\mathrm{T},\ \cdots\ \nabla_{n-1}\phi_{n-1}(x,y)^\mathrm{T},\ \nabla_n \tilde{\phi}_n(x,y)^\mathrm{T},\ 0,\ 0,\ \cdots\right]^\mathrm{T}. \tag{6}$$

With these definitions, the first $d(n+1)$ elements of the vector $\nabla U_n(x,y)$ are the partial derivatives of $U_n(x_0, \ldots, x_n, y_0, \ldots, y_n)$ with respect to $(x_0, \ldots, x_n)$, whilst the other elements of the vector $\nabla U_n(x,y)$ are zero. Treatment of $\nabla U_n(x,y)$ as an infinitely long vector allows us to consider $(\nabla U_n(\cdot, y))_{n \geq 0}$ as a sequence of vector fields on the common Hilbert space $l_2(\gamma)$. Define also

$$\alpha_{\gamma,n}(y) := \sum_{m=0}^n \gamma^{n-m}\beta_m(y), \quad \beta_m(y) := \|\nabla_m \phi_m(0,y)\|^2 \vee \|\nabla_m \tilde{\phi}_m(0,y)\|^2, \tag{7}$$

$$\eta_n(r,y) := \sup_{\|x\|_{\gamma,n}^2 \leq r} \|\nabla_n \phi_n(x,y)\|^2 \vee \|\nabla_n \tilde{\phi}_n(x,y)\|^2. \tag{8}$$

3 Quantitative $l_2(\gamma)$ bounds for maximum a-posteriori estimators

Condition 1.

a) μ, f, and $g(\cdot, y)$ for all $y \in \mathbb{Y}$, are everywhere strictly positive and continuously differentiable.

b) there exist constants $\zeta, \tilde{\zeta}, \theta$ such that $0 \leq \theta < \zeta/2 \wedge \tilde{\zeta}$, and for all $x, x' \in \mathbb{R}^\mathbb{N}$ and all $y \in \mathbb{Y}^\mathbb{N}$,

$$\begin{aligned} &\langle x_n - x_n', \nabla_n \phi_n(x,y) - \nabla_n \phi_n(x',y)\rangle \\ &\leq -\zeta\|x_n - x_n'\|^2 + \theta\|x_n - x_n'\|\left(\|x_{n-1} - x_{n-1}'\| + \|x_{n+1} - x_{n+1}'\|\right), \quad \forall n \geq 1, \\ &\left\langle x_n - x_n', \nabla_n \tilde{\phi}_n(x,y) - \nabla_n \tilde{\phi}_n(x',y)\right\rangle \\ &\leq \begin{cases} -\tilde{\zeta}\|x_0 - x_0'\|^2 + \theta\|x_0 - x_0'\|\|x_1 - x_1'\|, & n = 0, \\ -\tilde{\zeta}\|x_n - x_n'\|^2 + \theta\|x_n - x_n'\|\|x_{n-1}' - x_{n-1}'\|, & \forall n \geq 1. \end{cases} \end{aligned}$$

Theorem 1. *Assume that Condition 1 holds, and with $\zeta, \tilde{\zeta}, \theta$ as therein, let γ be any value in $(0,1]$ such that:*

$$\zeta > \theta\frac{(1+\gamma)^2}{2\gamma}, \qquad \tilde{\zeta} > \theta\frac{(1+\gamma)}{2\gamma}. \tag{9}$$

Then with any λ such that:

$$0 < \lambda \leq \left\{\zeta - \theta\frac{(1+\gamma)^2}{2\gamma}\right\} \wedge \left\{\tilde{\zeta} - \theta\frac{(1+\gamma)}{2\gamma}\right\}, \tag{10}$$

and any $n \geq 0$,

$$\langle x - x', \nabla U_n(x,y) - \nabla U_n(x',y)\rangle_\gamma \geq \lambda\|x - x'\|_\gamma^2, \quad \forall x, x' \in l_2^n(\gamma), y \in \mathbb{Y}^\mathbb{N}. \tag{11}$$

For any $y \in \mathbb{Y}^{\mathbb{N}}$, amongst all the vectors in $l_2^n(\gamma)$ there is a unique vector ξ_n such that $\nabla U_n(\xi_n, y) = 0$, and

$$\sup_{m\in\mathbb{N}_0, m\geq n} \|\xi_n - \xi_m\|_\gamma^2 \leq \frac{\gamma^n}{\lambda^2}\eta_n\left(\lambda^{-2}\alpha_{\gamma,n}(y), y\right) + \frac{\gamma^{n+1}}{\lambda^2}\eta_{n+1}\left(\gamma\lambda^{-2}\alpha_{\gamma,n}(y), y\right) + \frac{1}{\lambda^2}\sum_{k=n+2}^{\infty}\gamma^k\beta_k(y). \quad (12)$$

The proof of Theorem 1 is in appendix B. We pass several remarks on this result.

- To connect Theorem 1 with the setup from section 1, note that since $\nabla U_n(\xi_n, y) = 0$, the first $d(n+1)$ elements of the vector ξ_n solve the estimation problem (2), and since $\xi_n \in l_2^n(\gamma)$, the remaining elements of ξ_n are zero. Note that the dependence of ξ_n on y is not shown in the notation.
- When Condition 1 holds, there always exists $\gamma \in (0,1)$ satisfying (9) and λ satisfying (10) because $\lim_{\gamma\to 1}(1+\gamma)^2/2\gamma = \lim_{\gamma\to 1}(1+\gamma)/\gamma = 2$ and Condition 1 requires $0 \leq \theta < \zeta/2 \wedge \tilde{\zeta}$.
- If the right hand side of (12) converges to zero as $n \to \infty$, then $(\xi_n)_{n\geq 0}$ is a Cauchy sequence in $l_2(\gamma)$. Whether or not this convergence occurs depends on value of γ, the ingredients μ, f and g of the HMM and the y in question, and we shall return this topic in section 4.

From hereon (11) will be referred to as "decay-convexity" of U_n. Note that when $\gamma = 1$, (11) says exactly that $p(x_0, \ldots, x_n|y_0, \ldots, y_n)$ is λ-strongly log-concave in $(x_0, \ldots, x_n)$ in the sense of [18], which guarantees that $p(x_0, \ldots, x_n|y_0, \ldots, y_n)$ has a unique maximiser. This begs the question, if $(\xi_{n,0}, \cdots, \xi_{n,n})$ is a maximiser of $p(x_0, \ldots, x_n|y_0, \ldots, y_n)$, then can ξ_∞ be characterized as a maximiser of some function? As we shall see next, the answer is to this question appears to be "no", but an alternative and closely related interpretation of ξ_∞ is available if one introduces the vector field $\partial U : \mathbb{R}^{\mathbb{N}} \times \mathbb{Y}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$,

$$\partial U(x, y) := -[\nabla_0\tilde{\phi}_0(x, y)^{\mathrm{T}},\ \nabla_1\phi_1(x, y)^{\mathrm{T}},\ \nabla_2\phi_2(x, y)^{\mathrm{T}},\ \cdots]^{\mathrm{T}}. \quad (13)$$

Condition 2. For some given $y \in \mathbb{Y}^{\mathbb{N}}$,
a) there exists a finite constant χ such that for all n and $x \in l_2(\gamma)$,

$$\|\nabla_n\phi_n(x, y)\|^2 \vee \|\nabla_n\tilde{\phi}_n(x, y)\|^2 \leq \beta_n(y) + \chi\left(\|x_{n-1}\|^2 + \|x_n\|^2 + \|x_{n+1}\|^2\right),$$

b) $\sum_{n=0}^{\infty}\gamma^n\beta_n(y) < \infty$,
c) $x \mapsto \partial U(x, y)$ is continuous in $l_2(\gamma)$.

Theorem 2 below shows that, under the assumptions of Theorem 1 and additionally Condition 2, ξ_∞ is the fixed point of an infinite dimensional ODE, associated with the vector field $\partial U(\cdot, y)$. It is important to note here that the vector $\partial U(x, y) \in \mathbb{R}^{\mathbb{N}}$ is the element-wise limit as $n \to \infty$ of the vector $\nabla U_n(x, y)$. Indeed it can be read off from (6) that for any given x and y, each element of the vector $\nabla U_n(x, y)$ is constant in n for all n large enough. However, in general for some given x, y, the limit $\lim_{n\to\infty} U_n(x, y)$ does not exist, so it is a characterization of ξ_∞, as a fixed point of an ODE, rather than a maximiser of some function which is fruitful.

Theorem 2. *In addition to the assumptions of Theorem 1 and with γ as therein, assume that for some given $y \in \mathbb{Y}^{\mathbb{N}}$ Condition 2 holds. Then with λ as in Theorem 1,*

$$\langle x - x', \partial U(x, y) - \partial U(x', y)\rangle_\gamma \geq \lambda\|x - x'\|_\gamma^2, \qquad \text{for all } x, x' \in l_2(\gamma). \quad (14)$$

There exists a globally defined and unique flow $\Phi : (t, x) \in \mathbb{R}_+ \times l_2(\gamma) \mapsto \Phi(t, x) \in l_2(\gamma)$ which solves the Fréchet ordinary differential equation,

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi(t, x) = -\partial U(\Phi(t, x), y), \qquad \Phi(0, x) = x, \quad (15)$$

this flow has a unique fixed point, $\xi_\infty \in l_2(\gamma)$, and this point satisfies $\partial U(\xi_\infty, y) = 0$. With $(\xi_n)_{n\geq 0}$ as in Theorem 1,

$$\sup_{m\in\mathbb{N}_0\cup\{\infty\}, m\geq n} \|\xi_n - \xi_m\|_\gamma^2 \leq \frac{1}{\lambda^2}\left(\gamma^{n-1}\alpha_{\gamma,n}(y)\frac{2\chi}{\lambda^2} + \sum_{k=n}^{\infty}\gamma^k\beta_k(y)\right), \quad (16)$$

and $\|\xi_n - \xi_\infty\|_\gamma \to 0$ as $n \to \infty$.

The proof of Theorem 2 is in appendix B. In summary, the assumptions a)-b) of Condition 2 ensure that $\partial U(\cdot, y)$ maps $l_2(\gamma)$ to itself. Combined with the continuity in assumption c) in Condition 2 and (14), this allows an existence and uniqueness result of [4] for dissipative ordinary differential equations on Banach spaces to be applied. It is from here that the Fréchet derivative (15) arises. Background information about Fréchet derivatives is given in appendix A.1.

4 Discussion

4.1 A roadmap for applying Theorems 1 and 2

Our next objective is to set out and demonstrate mathematical tools which enable application of Theorems 1 or 2. For Theorem 1 there are two main tasks in such an application: firstly verifying Condition 1 and secondly proving that the right-hand side of (12) converges to zero as $n \to \infty$. Addressing these two tasks separately allows us to accommodate naturally the case of misspecified HMMs, in which the observation sequence $y = (y_n)_{n \geq 0}$ appearing in (12) need not be a realization from the HMM specified by μ, f, g, nor in fact any HMM. In sections 4.2 and 4.3 we give intermediate and generic results to help tackle these two main tasks for classes of HMMs, with calculations for a specific example in section 4.4.

In order to apply Theorem 2, one must additionally verify Condition 2, and calculations for a specific example are given in 4.5. In section 4.6 we address the use of Theorem 2 to quantify the error associated with parallelized approximation to MAP estimators, demonstrated with numerical results in the context of a model of neural population activity in section 4.7.

4.2 Verifying Condition 1

Lemma 1 focuses on a class of models with linear and Gaussian signals together with conditional distributions of the observations given signals which are not necessarily log-concave as a function of the signal. By contrast, Lemma 2 addresses a class of nonlinear, non-Gaussian signal models, combined with conditional distributions of observations given signal which are strongly log-concave. Here the minimum and maximum eigenvalues of a real, symmetric matrix, say B, are denoted $\rho_{\min}(B)$, $\rho_{\max}(B)$.

Lemma 1. *Assume a) and b):*
a) The signal satisfies the following vector autoregressive model:

$$X_n = AX_{n-1} + b + W_n, \tag{17}$$

where for $n \in \mathbb{N}$, $W_n \sim \mathcal{N}(0, \Sigma)$ is independent of other random variables, $X_0 \sim \mathcal{N}(b_0, \Sigma_0)$, Σ and Σ_0 are positive definite, A is a $d \times d$ matrix and b and b_0 are length-d vectors.
b) For each $y \in \mathbb{Y}$, $g(\cdot, y)$ is strictly positive, continuously differentiable and there exists $\lambda_g \in \mathbb{R}$ such that for all $x, x' \in \mathbb{R}^d$, $y \in \mathbb{Y}$,

$$\langle x - x', \nabla_x \log g(x, y) - \nabla_x \log g(x', y) \rangle \leq \lambda_g \|x - x'\|^2. \tag{18}$$

If the inequality $\theta < \zeta/2 \wedge \tilde{\zeta}$ is satisfied by:

$$\zeta = \frac{1 + \rho_{\min}(A^{\mathrm{T}} A)}{\rho_{\max}(\Sigma)} - \lambda_g, \tag{19}$$

$$\tilde{\zeta} = \frac{1}{\rho_{\max}(\Sigma)} \wedge \left\{ \frac{1}{\rho_{\max}(\Sigma_0)} + \frac{\rho_{\min}(A^{\mathrm{T}} A)}{\rho_{\max}(\Sigma)} \right\} - \lambda_g, \tag{20}$$

$$\theta = \frac{\rho_{\max}(A^{\mathrm{T}} A)^{1/2}}{\rho_{\min}(\Sigma)}, \tag{21}$$

then Condition 1 holds.

The proof is in section appendix C. We pass the following remarks on this result:

- The condition (18) is called semi-log-concavity of $x \mapsto g(x, y)$, generalizing log-concavity by allowing $\lambda_g \in \mathbb{R}$, rather than only $\lambda_g \leq 0$.

- The condition $\theta < \zeta/2 \wedge \tilde{\zeta}$ can be interpreted as balancing the magnitude of temporal correlation in (17) against the fluctuations of W_n and the degree to which the mapping $x \mapsto g(x,y)$ is informative about x. As $\lambda_g \to -\infty$ the mapping $x \mapsto g(x,y)$ becomes more strongly log-concave, and by inspection of (19)-(21) the condition $\theta < \zeta/2 \wedge \tilde{\zeta}$ can always be achieved if λ_g takes a negative value large enough in magnitude, with other quantities on the right of the equations (19)-(21) held constant. On the other hand, if $\rho_{\max}(\Sigma)^{-1} \wedge \rho_{\max}(\Sigma_0)^{-1} > \lambda_g$, which implies $\zeta \wedge \tilde{\zeta} > 0$ for any value of $\rho_{\min}(A^{\mathrm{T}}A)$, the condition $\theta < \zeta/2 \wedge \tilde{\zeta}$ can be achieved if $\rho_{\max}(A^{\mathrm{T}}A)$ is small enough.
- The fact that ζ, $\tilde{\zeta}$ and θ in (19)-(21) depend only on eigenvalues of A, Σ and Σ_0 and the semi-concavity parameter λ_g means they, and consequently λ and γ, do not necessarily depend on dimension. As a simple example consider the case: $\lambda_g \leq 0$, $A = aI_d$ and $\Sigma = \sigma^2 I_d$, with $|a| < 1$ and $\sigma^2 > 0$. In this situation $\theta < \zeta/2 \wedge \tilde{\zeta}$ holds, and γ and λ can be chosen to depend only on $|a|$ and σ^2 and be such that (10) holds.

Lemma 2. *Assume:*
a) the signal satisfies

$$X_n = A(X_{n-1}) + W_n, \tag{22}$$

where for $n \in \mathbb{N}$, W_n is independent of other random variables and has density proportional to $e^{-\psi(w)}$, and $\mu(x) \propto e^{-\psi_0(x)}$, where $\psi, \psi_0 : \mathbb{R}^d \to \mathbb{R}$ are continuously differentiable and have bounded, Lipschitz gradients in the sense that:

$$\sup_w \|\nabla\psi_0(w)\| \vee \|\nabla\psi(w)\| \leq L_\psi,$$
$$\|\nabla\psi(w) - \nabla\psi(w')\| \vee \|\nabla\psi_0(w) - \nabla\psi_0(w')\| \leq L_{\nabla\psi}\|w - w'\|,$$

A is continuously differentiable, and has bounded, Lipschitz gradient in the sense that:

$$\sup_x \|\nabla A(x)\|_{\mathrm{op}} \leq L_A, \quad \|\nabla A(x) - \nabla A(x')\|_{\mathrm{op}} \leq L_{\nabla A}\|x - x'\|,$$

where $\nabla A(x)$ is the Jacobian matrix of $x \mapsto A(x)$, and $\|\cdot\|_{\mathrm{op}}$ is the Euclidean operator norm.
b) for each $y \in \mathbb{Y}$, $x \mapsto g(x,y)$ is strictly positive, continuously differentiable and there exists $\lambda_g < 0$ such that for all $x, x' \in \mathbb{R}^d$, $y \in \mathbb{Y}$,

$$\langle x - x', \nabla_x \log g(x,y) - \nabla_x \log g(x',y)\rangle \leq \lambda_g\|x - x'\|^2. \tag{23}$$

If $\theta < \zeta/2 \wedge \tilde{\zeta}$ is satisfied with:

$$\zeta = \tilde{\zeta} = -(L_{\nabla\psi} + L_A^2 L_{\nabla\psi} + L_\psi L_{\nabla A}) - \lambda_g,$$
$$\theta = L_{\nabla\psi} L_A,$$

then Condition 1 holds.

The proof is in section appendix C.

4.3 Behaviour of $\alpha_{\gamma,n}(y)$ and $\beta_n(y)$ as $n \to \infty$

Lemma 3 and Lemma 4 are intermediate results which will be applied to bound the right hand sides of (12) and (16). The proofs of these two lemmas are in appendix C.

Lemma 3. *For any nonnegative random variables $(Z_n)_{n\geq 0}$, if there exists $p > 0$ and $s > 0$ such that*

$$\sup_{n\geq 0} \frac{\mathbf{E}[Z_n^s]}{(n+1)^p} < \infty,$$

then for any $\rho \in (0,1)$,

$$\sup_{n\geq 0} \rho^n Z_n < \infty, \quad a.s.$$

Lemma 4. *If $Y = (Y_n)_{n\geq 0}$ is any $\mathbb{Y}$-valued stochastic process with law denoted $\mathbf{P}_Y$ such that for some $s \in (0,1]$ and $p > 0$*

$$\sup_{n\geq 0} \frac{1}{(n+1)^p} \mathbf{E}\left[\| \nabla_x \log g(x, Y_n)|_{x=0} \|^{2s}\right] < \infty,$$

then for any $\rho \in (0,1)$ and any $\gamma > 0$,

$$\sup_{n\geq 0} \rho^n \alpha_{\gamma,n}(y) < \infty \quad \textit{and} \quad \sup_{n\geq 0} \rho^n \beta_n(y) < \infty, \quad \textit{for } \mathbf{P}_Y\textit{-almost all } y.$$

For any $\tilde{\rho} \in (\rho, 1)$, there exist $C(y)$ and $D(y)$ such that $C(y) \vee D(y) < \infty$ for $\mathbf{P}_Y$-almost all y, and

$$\rho^n \alpha_{\gamma,n}(y) \leq \tilde{\rho}^n C(y), \quad \textit{and} \quad \sum_{k=n}^{\infty} \rho^k \beta_k(y) \leq \tilde{\rho}^n D(y), \quad \textit{for all } n \textit{ and } \mathbf{P}_Y\textit{-almost all } y.$$

4.4 An example with a Gaussian signal and heavy-tailed observations

In this section we consider an instance of the model class addressed in Lemma 1 and the application of Theorem 1 to it. With $d = 1$ and $\mathbb{Y} = \mathbb{R}$ we consider

$$X_n = AX_{n-1} + W_n, \qquad Y_n = X_n + V_n, \tag{24}$$

where $A \in \mathbb{R}$, $|A| < 1$, $W_n \sim \mathcal{N}(0,1)$, and V_n follows a Student's-t distribution with one degree of freedom:

$$g(x,y) = \pi^{-1/2}\{1 + (y-x)^2\}^{-1}. \tag{25}$$

The mapping $x \mapsto g(x,y)$ is not log-concave, so (18) is not satisfied with any $\lambda_g \leq 0$, but it is satisfied for some $\lambda_g > 0$. Indeed we have:

$$\frac{\partial}{\partial x} \log g(x,y) = \frac{-2(x-y)}{1+(x-y)^2}, \qquad \frac{\partial^2}{\partial x^2} \log g(x,y) = -2\frac{1-2(x-y)^2}{[1+(x-y)^2]^2},$$

from which follows that $\frac{\partial^2}{\partial x^2} \log g(x,y) \leq 2$, uniformly in x and y, and in turn that (18) holds with $\lambda_g := 2$.

For (24) with $d = 1$, the matrix A in the signal model (17) is reduced to a single real-valued number and $\rho_{\max}(A^{\mathrm{T}}A) = A^2$. Thus in accordance with the discussion following Lemma 1, if we take the scalar $|A| < 1$ small enough, then $\theta < \zeta/2 \wedge \tilde{\zeta}$ is achieved and thus by Lemma 1, Condition 1 holds and Theorem 1 may be applied.

Our next step is to obtain a simplified bound for the right-hand side of the bound (12) from Theorem 1. By direct calculations we find:

$$\left[\frac{\partial}{\partial x} \log g(x,y)\bigg|_{x=0}\right]^2 \leq 4|y_n|^2, \tag{26}$$

there exists a finite constant χ such that:

$$|\nabla_n \phi_n(x)|^2 \vee |\nabla_n \tilde{\phi}_n(x)|^2 \leq \chi(|x_{n-1}|^2 + |x_n|^2 + |x_{n+1}|^2) + 4|y_n|^2$$

and

$$\beta_n(y) \leq 4|y_n|^2, \qquad \eta_n(r,y) \leq \frac{r}{\gamma} + 4|y_n|^2.$$

From (12) we obtain:

$$\sup_{m\in\mathbb{N}_0, m\geq n} \|\xi_n - \xi_m\|_\gamma^2 \leq \frac{\gamma^n}{\lambda^2}\left(\chi\lambda^{-2}\frac{\alpha_{\gamma,n}(y)}{\gamma} + 4|y_n|^2\right) + \frac{\gamma^{n+1}}{\lambda^2}\left(\chi\lambda^{-2}\alpha_{\gamma,n}(y) + 4|y_{n+1}|^2\right) + \frac{1}{\lambda^2}\sum_{k=n+2}^{\infty} \gamma^k \beta_k(y). \tag{27}$$

We now seek assumptions on the observations under which the right hand side of (27) converges to zero as $n \to \infty$, in which case $(\xi_n)_{n\geq 0}$ is a Cauchy sequence. We first consider the case of a well-specified model, where $\mathbf{P}_Y$ is the law of the random variables $(Y_n)_{n\geq 0}$ distributed according to (24). We approach this using Lemma 4 and Lemma 3.

From (24), using $|A| < 1$ and the distributional assumptions on $(W_n)_{n\geq 0}$ and $(V_n)_{n\geq 0}$, we have for any $s \in (0, 1/2)$, $(|a| + |b|)^{2s} \leq |a|^{2s} + |b|^{2s}$,

$$\sup_{n\geq 0} \mathbf{E}[|Y_n|^{2s}] \leq \sup_{n\geq 0} \mathbf{E}[|X_n|^{2s}] + \mathbf{E}[|V_0|^{2s}] < \infty,$$

and using (26)

$$\sup_{n\geq 0}\mathbf{E}\left[\left(\frac{\partial}{\partial x}\log g(x,Y_n)\Big|_{x=0}\right)^{2s}\right]\leq 4^s\sup_{n\geq 0}\mathbf{E}[|Y_n|^{2s}]<\infty. \tag{28}$$

Having established (28) we may apply Lemma 4 to bound $\alpha_{\gamma,n}(y)$ and $\sum_{k\geq n+1}\gamma^k\beta_k(y)$ and Lemma 3 with there $Z_n=|Y_n|^2$ to bound $\gamma^n|Y_n|^2$. We thus find that for any $\tilde{\gamma}\in(\gamma,1)$ there exists $C(y)$ such that $C(y)<\infty$ for $\mathbf{P}_Y$-almost all y, and the right hand side of (27) is bounded by $C(y)\tilde{\gamma}^n$ for $\mathbf{P}_Y$-almost all y. This implies that there exists a set in $\mathbb{Y}^\star_\gamma\subset\mathbb{Y}^{\mathbb{N}}$ such that under (24), $\mathbf{P}_Y(\mathbb{Y}^\star_\gamma)=1$, for any $y\in\mathcal{Y}$, the right hand side of (27) converges to zero as $n\to\infty$, and ξ_n is thus a Cauchy sequence.

Concerning the misspecified case, suppose that $Y=(Y_n)_{n\geq 0}$ does not necessarily follow (24) but rather is any $\mathbb{R}$-valued stochastic process, with law denoted $\mathbf{P}_Y$, such that for some $p>0$ and $s>0$:

$$\sup_{n\geq 0}\frac{\mathbf{E}[|Y_n|^{2s}]}{(1+n)^p}<\infty.$$

Then by Lemma 3, $\sup_{n\geq 0}\gamma^n|y_n|^2<\infty$, for $\mathbf{P}_Y$-almost all y; via (26),

$$\sup_{n\geq 0}\mathbf{E}\left[\left(\frac{\partial}{\partial x}\log g(x,Y_n)\Big|_{x=0}\right)^{2s}\right]<\infty,$$

and Lemma 4 can be applied to again show that the right hand side of (27) is a bounded by $C(y)\tilde{\gamma}^n$ for all $n\geq 0$, $\mathbf{P}_Y$-almost all y, $\tilde{\gamma}\in(\gamma,1)$ and some $C(y)<\infty$.

4.5 An example with a nonlinear, nonstationary, non-Gaussian signal

In this section we consider an instance of the model class addressed in Lemma 2 and the application of Theorem 2 to it. With $\mathbb{Y}=\mathbb{R}^d$, the model of interest is:

$$X_n=A(X_{n-1})+W_n \qquad Y_n=X_n+V_n, \tag{29}$$

where A and $(W_n)_{n\geq 0}$ are taken to satisfy the assumptions of Lemma 2 and $V_n\sim\mathcal{N}(0,\sigma_y^2 I_n)$, so that in Lemma 2 we can take $\lambda_g=-\sigma_y^{-2}$.

In the case $d=1$, an example of a function ψ (or ψ_0) which is continuously differentiable and has a bounded, Lipschitz gradient in the sense of Lemma 2 is the Huber function, with some $c>0$:

$$\psi(x)=\begin{cases}\frac{1}{2c}x^2, & |x|\leq c\\ |x|-\frac{c}{2}, & |x|>c.\end{cases}$$

By taking σ_y^2 small enough the condition $\theta<\zeta/2\wedge\tilde{\zeta}$ in Lemma 2 is satisfied and therefore Condition 1 holds. We now turn to verification of Condition 2. By direct calculation:

$$\beta_n(y)=\|\nabla_x\log g(x,y_n)|_{x=0}\|^2=\|y_n\|^2/\sigma_y^4. \tag{30}$$

and, using the same estimates as in the proof of Lemma 2, there exists a finite constant c such that for any $y\in\mathbb{Y}$,

$$\begin{aligned}\|\nabla_n\phi_n(x,y)-\nabla_n\phi_n(x',y)\|\vee\|\nabla_n\tilde{\phi}_n(x,y)-\nabla_n\tilde{\phi}_n(x',y)\|\\ \leq c\left(\|x_{n-1}-x'_{n-1}\|+\|x_n-x'_n\|+\|x_{n+1}-x'_{n+1}\|\right),\end{aligned} \tag{31}$$

and assumption a) of Condition 2 holds for all $y\in\mathbb{Y}^{\mathbb{N}}$.

Now let $\mathbb{Y}^\star_\gamma:=\{y=(y_0,y_1,\ldots)\in\mathbb{Y}^{\mathbb{N}}:\sum_n\gamma^n\|y_n\|^2<\infty\}$. Considering (30), clearly assumption b) of Condition 2 holds for any $y\in\mathbb{Y}^\star_\gamma$. It remains to verify assumption c) of Condition 2. To do so we combine the following identity:

$$\begin{aligned}\|\partial U(x,y)-\partial U(x',y)\|_\gamma^2=&\|\nabla_0\tilde{\phi}_0(x,y)-\nabla_0\tilde{\phi}_0(x',y)\|^2\\ &+\sum_{n\geq 1}\gamma^n\|\nabla_n\phi_n(x,y)-\nabla_n\phi_n(x',y)\|^2\end{aligned}$$

with (31).

We have thus established that Condition 2 holds and hence Theorem 2 holds for any $y \in \mathbb{Y}^\star_\gamma$. Our next objective is to exhibit conditions under which $\mathbf{P}_Y(\mathbb{Y}^\star_\gamma) = 1$ and establish the rate at which the right hand side of the bound (16) from Theorem 2 converges to zero exponentially hast as $n \to \infty$, for $\mathbf{P}_Y$-almost all y.

Consider the well-specified case, that is where $\mathbf{P}_Y$ is the law of the observations $(Y_n)_{n\geq 0}$ corresponding to (29). To control the growth of moments of $\|Y_n\|$ over time and apply Lemma 3 we shall assume that $\|A(x) - A(x')\| \leq \|x - x'\|$, in which case $\|A(x)\| \leq \|A(0)\| + \|x\|$,

$$\|X_n\| \leq \|A(X_{n-1})\| + \|W_n\| \leq \|A(0)\| + \|X_{n-1}\| + \|W_n\| \leq n\|A(0)\| + \sum_{m=1}^{n} \|W_n\| + \|X_0\|$$

and for any $s \in (1/2, 1)$, by Minkowski's inequality,

$$\begin{aligned}\mathbf{E}[\|Y_n\|^{2s}]^{1/2s} &\leq \mathbf{E}[\|X_n\|^{2s}]^{1/2s} + \mathbf{E}[\|V_0\|^{2s}]^{1/2s} \\ &\leq n\|A(0)\| + n\mathbf{E}[\|W_1\|^{2s}]^{1/2s} + \mathbf{E}[\|X_0\|^{2s}]^{1/2s} + \mathbf{E}[\|V_0\|^{2s}]^{1/2s}. \end{aligned} \tag{32}$$

This estimate allows us to apply Lemma 3 with Z_n there taken to be $\|Y_n\|^2$, and combined with (30) we obtain for any $\tilde{\gamma} \in (\gamma, 1)$,

$$\begin{aligned}\sum_{n=0}^{\infty} \gamma^n \beta_n(y) &\leq \sigma_y^{-4} \sum_{n=0}^{\infty} \gamma^n \|y_n\|^2 \\ &\leq \sigma_y^{-4} \left(\sup_{n\geq 0} \tilde{\gamma}^n \|y_n\|^2 \right) \sum_{n=0}^{\infty} \frac{\gamma^n}{\tilde{\gamma}^n} < \infty, \quad \text{for } \mathbf{P}_Y\text{-almost all } y.\end{aligned}$$

Therefore under (29), $\mathbf{P}_Y(\mathbb{Y}^\star_\gamma) = 1$. The bound (32) also allows Lemma 4 to be applied, and we conclude that there exists $C(y)$ such that $C(y) < \infty$ for $\mathbf{P}_Y$-almost all y, and the right hand side of (16) is bounded by $C(y)\tilde{\gamma}^n$ for all $n \geq 0$ and $\mathbf{P}_Y$-almost all y.

Again using Lemma 4, we reach the same conclusion in the misspecified case if $\mathbf{P}_Y$ is the law of any $\mathbb{R}^d$-valued stochastic process $Y = (Y_n)_{n\geq 0}$ such that for some $p > 0$ and $s > 0$,

$$\sup_{n\geq 0} \frac{\mathbf{E}[\|Y_n\|^s]}{(1+n)^p} < \infty.$$

4.6 Parallelized approximate optimization

The existence of the limit in (3), as has been verified via Theorems 1 and 2 in the preceeding sections, suggests that (2) can be solved approximately using a collection of optimization algorithms which process data segments in parallel. With Δ and $\ell = (n+1)/\Delta$ assumed to be integers, consider the partition:

$$\{0, \ldots, n\} = \bigcup_{k=1}^{\ell} A_k, \qquad A_k \coloneqq \{(k-1)\Delta, \ldots, k\Delta - 1\},$$

and for an integer $\delta > 0$ consider the δ-enlargement of each A_k,

$$A_k(\delta) \coloneqq \{m \in \{0, \ldots, n\} : \exists a \in A_k : |m - a| \leq \delta\}. \tag{33}$$

Suppose the ℓ optimization problems:

$$\operatorname*{arg\,max}_{x_{A_k(\delta)}} \; p(x_{A_k(\delta)} | y_{A_k(\delta)}), \qquad k = 1, \ldots, \ell, \tag{34}$$

where $x_{A_k(\delta)} = (x_m; m \in A_k(\delta))$, $y_{A_k(\delta)} = (y_m; m \in A_k(\delta))$, are solved in parallel. Then in a post-processing step, for each k, the components indexed by $A_k(\delta) \backslash A_k$ of the solution to $\arg\max_{x_{A_k(\delta)}} \; p(x_{A_k(\delta)} | y_{A_k(\delta)})$ are discarded, and what remains concatenated across k to give an approximation to the solution of (2). If it takes $T(n)$ time to solve (2) the speed-up from parallelization could be as much as a factor of $T(n)/T(\Delta + 2\delta)$. The Euclidean norm of the approximation error associated with the first segment in the parallelization scheme can be bounded using (12) or (16). The following is an immediate corollary of (16).

Corollary 1. *If the assumptions of Theorem 2 hold for some given $y \in \mathbb{Y}^{\mathbb{N}}$,*

$$\sup_{n \geq \Delta+\delta} \sum_{m=0}^{\Delta} \|\xi_{\Delta+\delta,m} - \xi_{n,m}\|^2 \leq \gamma^{-\Delta} \sup_{n \geq \Delta+\delta} \|\xi_{\Delta+\delta} - \xi_n\|_\gamma^2$$
$$\leq \frac{1}{\lambda^2} \left(\gamma^{\delta-1} \frac{2\chi}{\lambda^2} \alpha_{\gamma,\Delta+\delta}(y) + \sum_{k=\delta}^{\infty} \gamma^k \beta_{\Delta+k}(y) \right). \tag{35}$$

The right hand side of this bound can be controlled using the same tools and techniques demonstrated in sections 4.4 and 4.5 to analyse its convergence to zero as $\delta \to \infty$.

4.7 Application to a model of neural population activity

State-space models are used in neuroscience to examine time-varying dependence in the firing activity of neural populations and have been advocated for use in detecting cell assemblies in the brain – ensembles of neurons exhibiting coordinated firing – thought to play a key role in memory formation and learning [19, 5]. Neural spiking data in the form of multivariate binary time series are commonly modelled using random fields with time-dependent parameters. Here $y_n = (y_{n,k}^{(i)}, 1 \leq i \leq N, 1 \leq k \leq R) \in \{0,1\}^{NR} =: \mathbb{Y}$, where $y_{n,k}^{(i)} \in \{0,1\}$ indicates absence or presence of spiking activity of the i'th of N neurons during the n'th time bin of the k'th of R replicated experimental trials.

Similarly to [19] we consider a random field model for y_n given x_n, where the latent state has components $x_n = (x_n^{(i,j)}, 1 \leq i < j \leq N) \in \mathbb{R}^{N(N-1)/2}$, that is $d = N(N-1)/2.$, and:

$$g(x_n, y_n) = \exp\left\{ \sum_{j>i} \frac{x_n^{(i,j)}}{R} \sum_{k=1}^{R} (y_{n,k}^{(i)} - c^{(i)})(y_{n,k}^{(j)} - c^{(j)}) - C(x_n) \right\}. \tag{36}$$

Where $c^{(i)}$ is the average firing rate of the ith neuron over the R trials. The interest in this model is that the variables $x_n^{(i,j)}$ can be interpreted as a time-dependent statistical coupling between the firing activity of neurons i and j. The normalizing factor $C(x_n)$ is too expensive to compute for anything more than a handful of neurons and following [5] we consider the pseudo-likelihood approximation:

$$\tilde{g}(x_n, y_n) = \prod_{i=1}^{N} \prod_{k=1}^{R} \frac{\exp(y_{n,k}^{(i)} z_{n,k}^{(i)})}{1 + \exp(y_{n,k}^{(i)} z_{n,k}^{(i)})}, \tag{37}$$
$$z_{n,k}^{(i)} = \frac{1}{R} \left\{ \sum_{j<i} x_n^{(j,i)} (y_{n,k}^{(j)} - c^{(j)}) + \sum_{i<j} x_n^{(i,j)} (y_{n,k}^{(j)} - c^{(j)}) \right\}.$$

Both (36) and (37) are log-concave functions of x_n. Combining either with a prior model for the signal process as in (17) with $b = 0$, $A = aI_{N(N-1)/2}$, $|a| < 1$, $\Sigma = \sigma^2 I_{N(N-1)/2}$, where $I_{N(N-1)/2}$ is the identity matrix of size $N(N-1)/2$ and the prior distribution for X_0, μ, set to the stationary distribution of (17) hence by Lemma 1, Condition 1 holds and one may take γ, λ in Theorem 1 independently of d and hence N.

For both (36) and (37) the fact that $\mathbb{Y} = \{0,1\}^{NR}$ is a set with finitely many elements, combined with the linear-Gaussian nature of the signal implies that for any $y = (y_n)_{n \geq 0} \in \mathbb{Y}^{\mathbb{N}}$, $\sup_n \beta_n(y) < \infty$, $\sup_n \alpha_{n,\gamma}(y) < \infty$, and $\sup_n \eta_n(r, y) < \infty$ for any r. Therefore applied to either of these two models, for *any* $y \in \mathbb{Y}^{\mathbb{N}}$, the right hand side of the bound (12) from Theorem 1 converges to zero as $n \to \infty$.

4.8 Numerical results

We consider neural recordings of action potential spike trains from 30 medial prefrontal cortical neurons. The data were recorded using a 384-electrode Neuropixels probe [11] while an adult male rat navigated a 3-arm maze. Each recording was over a duration 35 seconds, from -15 to +20 seconds around the rat's arrival at particular location on the maze called the "reward point". The presence or absence of spiking per neuron was recorded in bins of width 10 millisecond, so that $n = 3500$. The data were divided in two subsets. The first consisted of $R = 76$ replicates of "correct" trials in which the rat navigated a particular route on the maze and consequently received a sweet reward at the reward point. The second data subset consisted of $R = 22$ "error" trials, in which the rat did not take that route and consequently

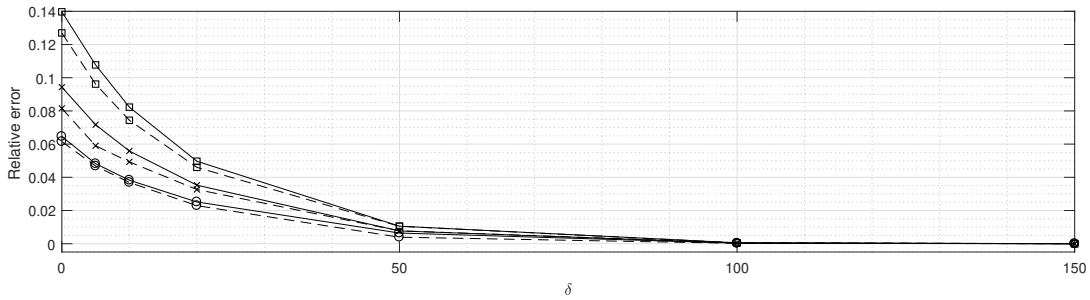

Figure 1: Relative error in the parallelization scheme vs. the overlap parameter δ. $N = 5$ neurons (dash), $N = 30$ neurons (solid), $\ell = 4$ cores ($\square$), $\ell = 8$ cores ($\times$) and $\ell = 6$ cores ($\circ$).

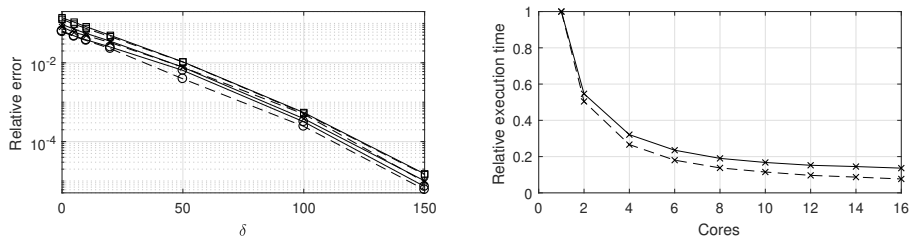

Figure 2: Left: relative error vs. overlap parameter δ as in figure 1 but on logarithmic scale. Right: Relative execution time vs. number of cores. $\delta = 100$ (solid), $\delta = 10$, (dashed).

received no reward. The scientific objective in analyzing these data is to study time-varying statistical interactions across the population of neurons as the rat approaches and passes through the reward point, and to identify differences in these interactions between the "correct" and "error" trials. We consider the pseudo-likelihood (37) combined with signal model described below that equation, and since there are $N = 30$ neurons we have $d = 435$.

Our first objective is to numerically evaluate the error and speed-up associated with the parallelization scheme. For this purpose, gradient-descent was used to approximately solve each of the ℓ optimization problems (34) for the neural model described above applied to the data subset of "correct" trials. The ℓ instances of gradient-descent were implemented in parallel using MATLAB's "parfor" command across ℓ Intel Sandy Bridge cores, each running at 2.6 Ghz, on a single blade of the University of Bristol's BlueCrystal Phase 3 cluster. 1.5×10^4 iterations of gradient descent were performed with a constant step size 10^{-8} in each instance. Let $\{\hat{\xi}_{n,0}(\ell,\delta), \ldots, \hat{\xi}_{n,n}(\ell,\delta)\}$ be the resulting approximation to (2) obtained by combining the approximate solutions to (34) as described in section 1. Figure 1 shows the relative error:

$$\frac{\sqrt{\sum_{m=0}^{n} \|\hat{\xi}_{n,m}(\ell,\delta) - \hat{\xi}_{n,m}(1,0)\|^2}}{\sqrt{\sum_{m=0}^{n} \|\hat{\xi}_{n,m}(1,0)\|^2}} \tag{38}$$

against the overlap parameter δ. Results are shown for the full data set from $N = 30$ neurons and also for a subset consisting of the first $N = 5$ neurons. Here Δ is determined by $\Delta = (n+1)/\ell$. The parameters of the state-space model were set to $a = 0.95$ and $\sigma = 10^{-4}$.

Corollary 1 suggests that (38) should decay to zero exponentially fast as δ grows. This is apparent in figure 1. A reduction in error as ℓ decreases can also be observed. The plot on the left of figure 2 shows the same results as figure 1 but with relative error on a logarithmic scale. Since the lines in figure 2 are close to parallel it appears that there is no degradation with N (hence d) in the exponential rate at which the relative error decays as δ grows. The plot on the right of figure 2 also shows the relative execution time of the parallelized scheme, that is the time taken to compute $\{\hat{\xi}_{n,m}(\ell,\delta), \ldots, \hat{\xi}_{n,n}(\ell,\delta)\}$ using $\ell = 2, 4, \ldots, 16$ cores divided by the time to compute $\{\hat{\xi}_{n,0}(1,0), \ldots, \hat{\xi}_{n,n}(1,0)\}$. Moving from 1 to 2 and 4 cores results in a roughly linear speed-up. Beyond 4 cores the speed-up is sublinear, which may be due to communication overhead associated with parallelization. From figure 2, $\delta = 100$ and $\ell = 4$ results in a relative error of less that 0.1% and a speed-up from parallelization of $1/0.32 = 3.125$.

Our next objective is to interpret the state-estimates obtained for the two data subsets. Figure 3 shows maximum a-posteriori state estimates of the pairwise coupling parameters x_n^{ij} for the two data

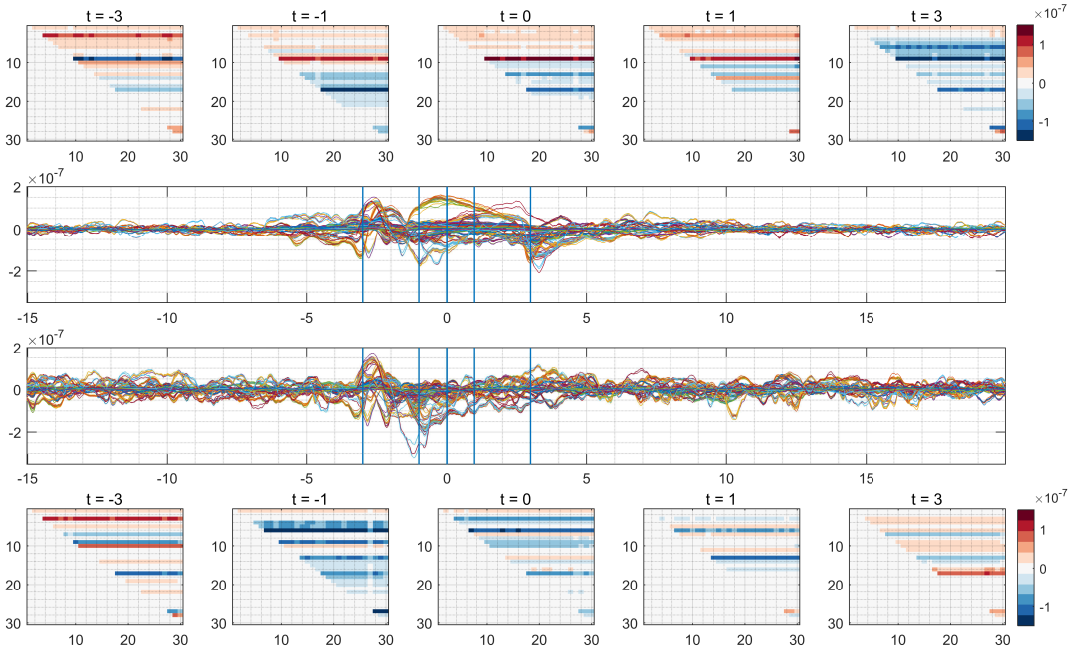

Figure 3: First and second rows: "correct" trials, third and fourth rows: "error" trials. Second and third row show $\{\hat{\xi}_{n,0}(4,100),\ldots,\hat{\xi}_{n,n}(4,100)\}$ with time on the horizontal axis in seconds relative to the reward delivery time. First and fourth rows display $\hat{\xi}_{n,m}(4,100)$ as a heat map, for m corresponding left-to-right to: 3 and 1 seconds before reward delivery, delivery time itself, and 1 and 3 seconds after reward delivery. These times are marked by the vertical blue lines in the second and third row plots.

subsets, consisting respectively of "correct" and "error" trials. The first and fourth rows in figure 3 display $\hat{\xi}_{n,m}(4,100)$ as a heat map, for five different time steps: m corresponding to 3 ($t=-3$) and 1 ($t=-1$) seconds before arrival at the reward point; at the reward point ($t=0$), and 1 ($t=1$) and 3 ($t=3$) seconds after pass the reward point. These times are marked by vertical blue lines in the second and third row plots in figure 3.

Signatures of neural population interaction emerge from the estimates of pairwise coupling parameters: on correct trials, the red coloring on the heat-maps corresponds to strong positive influence from a small minority of neurons onto a larger pool, at $t=-3$, $t=1$, $t=0$ and extending to $t=1$. Such anticipatory activity could reflect reward expectancy. Conversely, on "error" trials the predominantly blue colouring on the heatmaps at $t=-1$, $t=0$ and $t=1$ indicates negative pairwise interactions shortly before, during and after the reward point. The estimates on the "error" trials also present greater variation over time than in the "correct" trials, possibly reflecting an error signal or the consolidation of trial outcome-related information.

4.9 Comparison to the assumptions of [3]

The assumptions of [3, Thm 3.1] require that $x \mapsto \mu(x)$ and $(x,x') \mapsto f(x,x')$ are log-concave, and that $x \mapsto g(x,y)$ is strongly log-concave, uniformly in y. As per section 4.2, our Condition 1b) does not require all these conditions to hold simultaneously. Assumption (a4) (sic.) of [3, Thm 3.1] is that with $f(x,x') \propto e^{-\alpha(x,x')}$, there is a non-decreasing function $g:\mathbb{R}_+ \to \mathbb{R}_+$ growing to $+\infty$ not faster than polynomially, such that for all $M>0$,

$$\alpha(x,x') \le M \quad \Longrightarrow \quad \left|\frac{\partial^2}{\partial x \partial x'}\alpha(x,x')\right| \le g(M), \quad \text{for all } x,x'. \tag{39}$$

Putting aside the issue of once versus twice differentiability, this assumption is related to the terms multiplied by θ in our Condition 1b), but allows greater generality because $g(M)$ can grow with M, where as our Condition 1b) requires a value of θ uniform in x,x'. [3, Thm 3.1] also places an assumption on the asymptotic behaviour of $n^{-1}U_n(X_0,\ldots,X_n,Y_0,\ldots,Y_n)$ as $n\to\infty$ which may be regarded as a counterpart to the assumptions of Lemma 4.

A Fréchet derivatives and differential equations

A.1 Fréchet derivatives

The following definitions can be found in [10, App. A]. For Banach spaces V, W over $\mathbb{R}$, with respective norms $\|\cdot\|_V$, $\|\cdot\|_W$, a function $\varphi : V \to W$ has a directional derivative at $x \in V$ in direction $v \in V$ if there exists $\partial\varphi(v; x) \in W$ such that

$$\lim_{\epsilon\to 0}\left\|\frac{\varphi(x+\epsilon v)-\varphi(x)}{\epsilon} - \partial\varphi(v;x)\right\|_W = 0.$$

The function φ is Gâteaux differentiable at x if $\partial\varphi(v; x)$ exists for all $v \in V$ and $D\varphi(\cdot; x) : v \mapsto \partial\varphi(v; x)$ is a bounded linear operator from V to W, in which case $D\varphi(\cdot; x)$ is called the Gâteaux derivative at x. The function φ is additionally Fréchet differentiable at x if

$$\lim_{\epsilon\to 0}\sup_{v:\|v\|_V=1}\left\|\frac{\varphi(x+\epsilon v)-\varphi(x)}{\epsilon} - \partial\varphi(v;x)\right\|_W = 0, \tag{40}$$

in which case the operator $D\varphi(\cdot; x)$ is called the Fréchet derivative at x.

A.2 Ordinary differential equations on the Hilbert space

In the following proposition the operator of orthogonal projection from $l_2(\gamma)$ to $l_2^n(\gamma)$ is written Π_n.

Proposition 1. *For a given triple (γ, F, n) consisting of a constant $\gamma \in (0, 1]$, a mapping $F : l_2(\gamma) \to l_2(\gamma)$ and $n \geq 0 \cup \{\infty\}$, assume that a)-c) hold:*
a) F is continuous with respect to the norm $\|\cdot\|_\gamma$ on $l_2(\gamma)$,
b) there exists $\lambda > 0$ such that for all $x, x' \in l_2^n(\gamma)$,

$$\langle x - x', F(x) - F(x')\rangle_\gamma \leq -\lambda\|x - x'\|_\gamma^2,$$

c) $F(x) = F \circ \Pi_n(x)$ and $F(x) \in l_2^n(\gamma)$ for all $x \in l_2(\gamma)$.

Then there exists a globally defined and unique flow $\Phi : (t, x) \in \mathbb{R}_+ \times l_2(\gamma) \mapsto \Phi(t, x) \in l_2(\gamma)$ solving the Fréchet differential equation,

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi(t,x) = F(\Phi(t,x)), \qquad \Phi(0,x) = x.$$

This flow has a unique in $l_2^n(\gamma)$ fixed point, ξ, and this point satisfies $F(\xi) = 0$ and $\|\Phi(t, x) - \xi\|_\gamma \leq e^{-\lambda t}\|x - \xi\|_\gamma$ for all $x \in l_2^n(\gamma)$ and $t \geq 0$.

The proof is postponed.

The term $\frac{\mathrm{d}}{\mathrm{d}t}\Phi(t, x)$ in Proposition 1 is an application of the Fréchet derivative of $\Phi(t, x)$ with respect to t, that is in (40), V is $\mathbb{R}$ equipped with the Euclidean norm, W is the Hilbert space $l_2(\gamma)$, and φ is the map $t \mapsto \Phi(t, x)$, where in the latter the x argument is regarded as fixed. Similarly with x fixed, and denoting the Fréchet derivative of $t \mapsto \Phi(t, x)$ at t by $D\Phi(\cdot; t, x)$, the quantity $\frac{\mathrm{d}}{\mathrm{d}t}\Phi(t, x)$ is precisely $D\Phi(1; t, x)$. Thus in particular,

$$\lim_{\delta\searrow 0}\left\|\frac{\Phi(t+\delta,x)-\Phi(t,x)}{\delta} - \frac{\mathrm{d}}{\mathrm{d}t}\Phi(t,x)\right\|_\gamma = 0,$$

which, in general, is stronger than the element-wise convergence of $[\Phi(t + \delta, x) - \Phi(t, x)]/\delta$ to $\frac{\mathrm{d}}{\mathrm{d}t}\Phi(t, x)$.

The following lemma will be used in the proof of Proposition 1.

Lemma 5. *If a triple (γ, F, n) satisfies the assumptions of Proposition 1, then with Φ as therein and any $x, x' \in l_2(\gamma)$,*

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\Phi(t,x) - \Phi(t,x')\|_\gamma^2 = 2\left\langle \Phi(t,x) - \Phi(t,x'), F(\Phi(t,x)) - F(\Phi(t,x'))\right\rangle_\gamma. \tag{41}$$

Proof. In the case $n < \infty$, assumption c) of Proposition 1 implies that only the first $d(n+1)$ elements of the vector $\Phi(t, x)$ depend on t, and in that case the lemma can be proved by the chain rule of elementary

differential calculus. The following proof is valid for any $n \geq 0 \cup \{\infty\}$ and uses the chain rule of Fréchet differentiation.

Pick any $x, v \in l_2(\gamma)$, write them as $x = [x_0^{\mathrm{T}}\, x_1^{\mathrm{T}} \cdots]^{\mathrm{T}}$, $v = [v_0^{\mathrm{T}}\, v_1^{\mathrm{T}} \cdots]^{\mathrm{T}}$ with each $x_k, v_k \in \mathbb{R}^d$. The first step is to prove that the mapping $\varphi(x) = \|x\|_\gamma^2$ is Fréchet differentiable everywhere in $l_2(\gamma)$, with Fréchet derivative $D\varphi(v; x) = 2\langle v, x\rangle_\gamma$.

Consider the existence of directional derivatives. For $m \in \mathbb{N}$ let e_m denote the vector in $l_2(\gamma)$ whose mth entry is 1 and whose other entries are zero. The directional derivative $\partial\varphi(e_m; x)$ clearly exists.

We now need to check the existence of directional derivatives of φ in arbitrary directions in $l_2(\gamma)$. To do so we shall validate the following four equalities:

$$\lim_{\epsilon\to 0} \frac{\varphi(x+\epsilon v) - \varphi(x)}{\epsilon} = \lim_{\epsilon\to 0}\lim_{m\to\infty} \frac{\varphi(x+\epsilon\Pi_m(v)) - \varphi(x)}{\epsilon} \tag{42}$$

$$= \lim_{m\to\infty}\lim_{\epsilon\to 0} \frac{\varphi(x+\epsilon\Pi_m(v)) - \varphi(x)}{\epsilon} \tag{43}$$

$$= \lim_{m\to\infty} 2\sum_{k=0}^{m} \gamma^k \langle v_k, x_k\rangle \tag{44}$$

$$= 2\langle v, x\rangle_\gamma . \tag{45}$$

For (42), we have for any $\epsilon > 0$,

$$\begin{aligned} |\varphi(x+\epsilon\Pi_m(v)) - \varphi(x+\epsilon v)| &\leq \sum_{k\geq m+1} \gamma^k \left| \|x_k\|^2 - \|x_k + \epsilon v_k\|^2 \right| \\ &\leq 3\sum_{k\geq m+1} \gamma^k \|x_k\|^2 + \epsilon^2 \sum_{k\geq m+1} \gamma^k \|v_k\|^2 \\ &\to 0, \quad \text{as} \quad m\to\infty, \end{aligned} \tag{46}$$

where the convergence holds since x and v are members of $l_2(\gamma)$.

Let $\nabla_k\varphi(x)$ be the vector in $\mathbb{R}^d$ whose ith entry is the partial derivative of $\varphi(x)$ with respect to the ith element of x_k, that is $\nabla_k\varphi(x) = 2x_k$. Since $\varphi(x) = \sum_{k=0}^{\infty}\gamma^k\|x_k\|^2$, the directional derivative in direction $\Pi_m(v)$ at x is given by:

$$\partial\varphi(\Pi_m(v); x) = \lim_{\epsilon\to 0} \frac{\varphi(x+\epsilon\Pi_m(v)) - \varphi(x)}{\epsilon} = \sum_{k=0}^{m} \langle v_k, \nabla_k\varphi(x)\rangle = 2\sum_{k=0}^{m} \gamma^k \langle v_k, x_k\rangle . \tag{47}$$

Let us now check that the convergence in (47) is uniform in m in order to verify the equality in (43). By the mean value theorem of elementary differential calculus, for any $\epsilon > 0$ there exists $y^{m,\epsilon}$ on the line segment between x and $x + \epsilon\Pi_m(v)$ (so $y_k^{m,\epsilon} = x_k$ for $k > m$) such that

$$\begin{aligned} &\sup_m \left| \frac{\varphi(x+\epsilon\Pi_m(v)) - \varphi(x)}{\epsilon} - 2\langle \Pi_m(v), x\rangle_\gamma \right| \\ &= \sup_m \left| \frac{\sum_{k=0}^{m} \epsilon \langle v_k, \nabla_k\varphi(y^{m,\epsilon})\rangle}{\epsilon} - 2\langle \Pi_m(v), x\rangle_\gamma \right| \\ &= 2\sup_m \left| \langle \Pi_m(v), y^{m,\epsilon} - x\rangle_\gamma \right| \\ &\leq 2\sup_m \|\Pi_m(v)\|_\gamma \|y^{m,\epsilon} - x\|_\gamma \\ &\leq 2\|v\|_\gamma^2 \epsilon, \end{aligned} \tag{48}$$

so the convergence in (47) is indeed uniform in m. Therefore (43) holds.

For the two remaining equalities, (44) is already proved in (47), and (45) holds by Cauchy-Schwartz combined with the facts that $x, v \in l_2(\gamma)$ and that absolute convergence of a series in $\mathbb{R}$ implies its convergence.

We have established that the directional derivative of φ at an arbitrary x in an arbitrary direction v exists and is given by $2\langle v, x\rangle_\gamma$. To prove that φ is everywhere Gâteaux differentiable, we also need to show that for each x, $D\varphi(\cdot; x) : v \mapsto 2\langle v, x\rangle_\gamma$ is a bounded operator from $l_2(\gamma)$ to $\mathbb{R}$. This follows from Cauchy-Schwartz:

$$\sup_{v\neq 0} \frac{2|\langle v, x\rangle_\gamma|}{\|v\|_\gamma} \leq 2\|x\|_\gamma < +\infty, \quad \text{for all } x \in l_2(\gamma).$$

To prove that φ is Fréchet differentiable everywhere in $l_2(\gamma)$, it suffices, by [10, App A, Prop A.3], to check that $D\varphi(\cdot;x)$ is operator-norm continuous in x. This follows again by the Cauchy-Schwartz inequality:

$$\sup_{v\neq 0}\frac{2|\langle v,x\rangle_\gamma-\langle v,x'\rangle_\gamma|}{\|v\|_\gamma}=\sup_{v\neq 0}\frac{2|\langle v,x-x'\rangle_\gamma|}{\|v\|_\gamma}\leq 2\|x-y\|_\gamma,\quad \text{for all } x,x'\in l_2(\gamma).$$

We have proved that $\varphi(x)=\|x\|_\gamma^2$ is Fréchet differentiable everywhere in $l_2(\gamma)$, with Fréchet derivative in direction v given by $D\varphi(v;x)=2\langle v,x\rangle_\gamma$.

The proof is completed by an application of the chain rule of Fréchet differentiation:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|\Phi(t,x)-\Phi(t,x')\|_\gamma^2&=D\varphi\left(\frac{\mathrm{d}}{\mathrm{d}t}\{\Phi(t,x)-\Phi(t,x')\};\Phi(t,x)-\Phi(t,x')\right)\\&=2\left\langle F\{\Phi(t,x)\}-F\{\Phi(t,x')\},\Phi(t,x)-\Phi(t,x')\right\rangle_\gamma.\end{aligned}$$

□

Proof of Proposition 1. Let $n\geq 0\cup\{\infty\}$ be as in the statement of the proposition. Applying assumptions b) and c) of the proposition, we have for any $x,x'\in l_2(\gamma)$,

$$\begin{aligned}\langle F(x)-F(x'),x-x'\rangle_\gamma&=\langle F\circ\Pi_n(x)-F\circ\Pi_n(x'),\Pi_n(x)-\Pi_n(x')\rangle_\gamma\\&\quad+\langle F\circ\Pi_n(x)-F\circ\Pi_n(x'),x-\Pi_n(x)-x'+\Pi_n(x')\rangle_\gamma\\&\leq-\lambda\|\Pi_n(x)-\Pi_n(x')\|_\gamma^2+0\\&\leq 0.\end{aligned}$$

This global dissipation condition, combined with assumption a) of the proposition allows the application of [4, Thm 3.4, p.41] on the Hilbert space $l_2(\gamma)$ to give the existence and uniqueness of the globally defined flow as required.

Under assumption c) of the proposition, $F(x)\in l_2^n(\gamma)$ for any $x\in l_2(\gamma)$.Then, since $\Phi(t,x)=x+\int_0^t F(\Phi(s,x))\mathrm{d}s$,we find that if $x\in l_2^n(\gamma)$, then $\Phi(t,x)\in l_2^n(\gamma)$ for all $t>0$. Now fix any $x,x'\in l_2^n(\gamma)$ and define $a(t):=\|\Phi(t,x)-\Phi(t,x')\|_\gamma^2$. Lemma 5 combined with assumption b) of the proposition gives:

$$\frac{\mathrm{d}}{\mathrm{d}t}a(t)\leq-2\lambda a(t),$$

from which it follows that

$$a(t)\leq a(0)\exp(-2\lambda t).$$

We have thus proved that $\|\Phi(t,x)-\Phi(t,x')\|_\gamma\leq e^{-\lambda t}\|x-x'\|_\gamma$ for all $x,x'\in l_2^n(\gamma)$. An application of the Banach fixed point theorem to the restriction of Φ to the Hilbert space $l_2^n(\gamma)$ then gives the existence of the unique (in l_2^n) fixed point ξ. Since ξ is a fixed point we have for all $t>0$

$$\xi=\Phi(t,\xi)=\xi+\int_0^t F(\Phi(s,\xi))\mathrm{d}s=\xi+F(\xi)t,$$

which implies $F(\xi)=0$. □

B Proofs for section 3

In Lemma 6 and the proof of Theorem 1 below we shall need the following generalization of the inner-product $\langle\cdot,\cdot\rangle_\gamma$ and norm $\|\cdot\|_\gamma$, for $n\geq 0$,

$$\langle x,x'\rangle_{\gamma,n}=\sum_{m=0}^{\infty}\gamma^{|m-n|}\langle x_m,x'_m\rangle,\qquad \|x\|_{\gamma,n}=\langle x,x\rangle_{\gamma,n}^{1/2},\qquad x,x'\in l_2(\gamma).\tag{49}$$

Lemma 6. *Assume that Condition 1 holds, and with $\zeta,\tilde{\zeta},\theta$ as therein and $\gamma\in(0,1]$ assume the following inequalities hold:*

$$\zeta>\frac{\theta}{2\gamma}(1+\gamma)^2,\qquad \tilde{\zeta}>\frac{\theta}{2\gamma}(1+\gamma).\tag{50}$$

Then for any λ such that:

$$0 < \lambda \leq \left\{\tilde{\zeta} - \frac{\theta}{2\gamma}(1+\gamma)^2\right\} \wedge \left\{\zeta - \frac{\theta}{2\gamma}(1+\gamma)\right\}, \tag{51}$$

any $n \geq 0$ and $m = 0, \ldots, n$,

$$\langle x - x', \nabla U_n(x,y) - \nabla U_n(x',y)\rangle_{\gamma,m} \geq \lambda \|x - x'\|^2_{\gamma,m}, \quad \textit{for all} \quad x, x' \in l^n_2(\gamma).$$

Proof. For any $x, x' \in l^n_2(\gamma)$ we have:

$$\begin{aligned}
&\langle x - x', \nabla U_n(x,y) - \nabla U_n(x',y)\rangle_{\gamma,m} \\
&= -\gamma^{|m|}\left\langle x_0 - x'_0, \nabla_0\tilde{\phi}_0(x) - \nabla_0\tilde{\phi}_0(x')\right\rangle \\
&\quad - \sum_{k=1}^{n-1}\gamma^{|k-m|}\left\langle x_k - x'_k, \nabla_k\phi_k(x) - \nabla_k\phi_k(x')\right\rangle \\
&\quad - \gamma^{|n-m|}\left\langle x_n - x'_n, \nabla_n\tilde{\phi}_n(x,y) - \nabla_n\tilde{\phi}_n(x',y)\right\rangle
\end{aligned}$$

$$\begin{aligned}
&\geq \gamma^m\left\{\tilde{\zeta}\|x_0 - x'_0\|^2 - \theta\|x_0 - x'_0\|\|x_1 - x'_1\|\right\} \\
&\quad + \sum_{k=1}^{n-1}\gamma^{|k-m|}\left\{\zeta\|x_k - x'_k\|^2 - \theta\|x_k - x'_k\|\left(\|x_{k-1} - x'_{k-1}\| + \|x_{k+1} - x'_{k+1}\|\right)\right\} \\
&\quad + \gamma^{n-m}\left\{\tilde{\zeta}\|x_n - x'_n\|^2 - \theta\|x_n - x'_n\|\|x_{n-1} - x'_{n-1}\|\right\} \\
&\geq \gamma^m\left\{(\tilde{\zeta} - \frac{\theta}{2})\|x_0 - x'_0\|^2 - \frac{\theta}{2}\|x_1 - x'_1\|^2\right\} \\
&\quad + \sum_{k=1}^{n-1}\gamma^{|k-m|}\left\{(\zeta - \theta)\|x_k - x'_k\|^2 - \frac{\theta}{2}\|x_{k-1} - x'_{k-1}\|^2 - \frac{\theta}{2}\|x_{k+1} - x'_{k+1}\|^2\right\} \\
&\quad + \gamma^{n-m}\left\{(\tilde{\zeta} - \frac{\theta}{2})\|x_n - x'_n\|^2 - \frac{\theta}{2}\|x_{n-1} - x'_{n-1}\|^2\right\} \qquad (52) \\
&= \gamma^m\left\{(\tilde{\zeta} - \frac{\theta}{2})\|x_0 - x'_0\|^2 - \frac{\gamma^{|1-m|}}{\gamma^m}\frac{\theta}{2}\|x_0 - x'_0\|^2\right\} \\
&\quad \sum_{k=1}^{n-1}\gamma^{|k-m|}\left\{(\zeta - \theta)\|x_k - x'_k\|^2 - \frac{\gamma^{|k-1-m|}}{\gamma^{|k-m|}}\frac{\theta}{2}\|x_k - x'_k\|^2 - \frac{\gamma^{|k+1-m|}}{\gamma^{|k-m|}}\frac{\theta}{2}\|x_k - x'_k\|^2\right\} \\
&\quad \gamma^{n-m}\left\{(\tilde{\zeta} - \frac{\theta}{2})\|x_n - x'_n\|^2 - \frac{\gamma^{|n-1-m|}}{\gamma^{n-m}}\frac{\theta}{2}\|x_n - x'_n\|^2\right\} \\
&\geq \gamma^m\{\tilde{\zeta} - \frac{\theta}{2}(1+\gamma^{-1})\}\|x_0 - x'_0\|^2 \\
&\quad + \{\zeta - \theta - \frac{\theta}{2}(\gamma + \gamma^{-1})\}\sum_{k=1}^{n-1}\gamma^{|k-m|}\|x_k - x'_k\|^2 \\
&\quad + \{\tilde{\zeta} - \frac{\theta}{2}(1+\gamma^{-1})\}\gamma^{n-m}\|x_n - x'_n\|^2 \\
&\geq \lambda\sum_{k=0}^{n}\gamma^{|k-m|}\|x_k - x'_k\|^2, \qquad (53)
\end{aligned}$$

where the first equality and inequality are due to (6) and Condition 1b); the second inequality uses a re-arrangement of the terms in the summation and the fact that for any $a, b \in \mathbb{R}$, $2|a||b| \leq |a|^2 + |b|^2$; the third inequality uses $\frac{\gamma^{|1-m|}}{\gamma^m} \vee \frac{\gamma^{|n-1-m|}}{\gamma^{n-m}} \leq \frac{1}{\gamma}$ for $0 \leq m \leq n$ and $\frac{\gamma^{|k-1-m|}}{\gamma^{|k-m|}} + \frac{\gamma^{|k+1-m|}}{\gamma^{|k-m|}} \leq \frac{1}{\gamma} + \gamma$; and the final inequality holds under the conditions on $\lambda, \zeta, \tilde{\zeta}, \theta$ and γ given in (9) and (10). □

Proof of Theorem 1. Throughout the proof, $n \geq 0$ and $\gamma \in (0, 1]$ are fixed. Considering $(\gamma, -\nabla U_n, n)$, let us validate the assumptions of Proposition 1 in the order: c), then a), then b). Assumption c) of

Proposition 1 holds due to the definition of ∇U_n in (6). Validating assumption a) of Proposition 1 requires that if $x \to x'$ in $l_2(\gamma)$ then $\|\nabla U_n(x,y) - \nabla U_n(x',y)\|_\gamma \to 0$. But we have already validated assumption c) of Proposition 1, so ∇U_n maps $l_2(\gamma)$ into $l_2^n(\gamma)$, and

$$\begin{aligned}\|\nabla U_n(x,y) - \nabla U_n(x',y)\|_\gamma^2 &= \|\nabla_0\tilde{\phi}_0(x,y) - \nabla_0\tilde{\phi}_0(x',y)\|^2 \\ &\quad + \sum_{m=1}^{n-1}\gamma^m\|\nabla_m\phi_m(x,y) - \nabla_m\phi_m(x',y)\|^2 \\ &\quad + \gamma^n\|\nabla_n\tilde{\phi}_n(x,y) - \nabla_n\tilde{\phi}_n(x',y)\|^2.\end{aligned}$$

Also by assumption c) of Proposition 1, $\nabla U_n(x,y)$ depends on x only through $(x_0,\ldots,x_n)$. These observations together with Condition 1a) validate assumption a) of Proposition 1. Assumption b) of Proposition 1 holds by an application of Lemma 6. This completes the verification of the assumptions of Proposition 1 for $(\gamma, -\nabla U_n, n)$ and thus establishes the existence of the fixed point ξ_n.

Our next step is to obtain bounds on $\|\xi_n\|_{\gamma,n}^2$ and $\|\xi_n\|_{\gamma,n+1}^2$. An application of Lemma 6 and Cauchy-Schwartz gives:

$$\|\xi_n\|_{\gamma,n}^2 \le \frac{1}{\lambda}\langle 0 - \xi_n, \nabla U_n(0,y) - 0\rangle_{\gamma,n} \le \frac{1}{\lambda}\|\xi_n\|_{\gamma,n}\|\nabla U_n(0,y)\|_{\gamma,n},$$

hence

$$\|\xi_n\|_{\gamma,n}^2 \le \frac{\alpha_{\gamma,n}(y)}{\lambda^2} \quad \text{and} \quad \|\xi_n\|_{\gamma,n+1}^2 = \gamma\|\xi_n\|_{\gamma,n}^2 \le \gamma\frac{\alpha_{\gamma,n}(y)}{\lambda^2}, \tag{54}$$

where the equality uses the fact that $\|\xi_{n,m}\| = 0$ for $m > n$.

Now fix any $m > n$. An application of Lemma 6 and Cauchy-Schwartz gives:

$$\|\xi_n - \xi_m\|_{\gamma,0}^2 \le \frac{1}{\lambda}\langle \xi_n - \xi_m, \nabla U_m(\xi_n,y) - 0\rangle_{\gamma,0} \le \frac{1}{\lambda}\|\xi_n - \xi_m\|_{\gamma,0}\|\nabla U_m(\xi_n,y)\|_{\gamma,0}. \tag{55}$$

Observe $\nabla U_n(\xi_n,y) = 0$ implies $\nabla_0\tilde{\phi}_0(\xi_n,y) = \nabla_k\phi_k(\xi_n,y) = 0$ for all $k < n$. Combining this fact with $\xi_n \in l_2^n(\gamma)$,(6), (8) and the bound (54) gives:

$$\begin{aligned}\|\nabla U_m(\xi_n,y)\|_{\gamma,0}^2 &= \sum_{k=n}^{m-1}\gamma^k\|\nabla\phi_k(\xi_n,y)\|^2 + \gamma^m\|\nabla\tilde{\phi}_m(\xi_n,y)\|^2 \\ &\le \sum_{k=n}^{n+1}\gamma^k\|\nabla\phi_k(\xi_n,y)\|^2 + \sum_{k=n+2}^{\infty}\gamma^k\beta_k(y) \\ &\le \gamma^n\eta_n\left(\frac{\alpha_{\gamma,n}(y)}{\lambda^2}, y\right) + \gamma^{n+1}\eta_{n+1}\left(\gamma\frac{\alpha_{\gamma,n}(y)}{\lambda^2}, y\right) + \sum_{k=n+2}^{\infty}\gamma^k\beta_k(y).\end{aligned} \tag{56}$$

The proof of the theorem is completed by combining this bound with (55). □

Proof of Theorem 2. The first step is to apply Proposition 1 to $(\gamma, -\partial U(\cdot,y), \infty)$. From its definition (13) combined with assumptions b) and c) of Condition 2, it is clear that $\partial U(\cdot,y)$ maps $l_2(\gamma)$ into itself and $\Pi_\infty = \mathrm{Id}$ by definition, so assumption c) of Proposition 1 is satisfied. Assumption c) of Condition 2 is exactly what is required for assumption a) of Proposition 1 to hold. Let us now verify assumption b) of Proposition 1. For any$x, x' \in l_2(\gamma)$ and $n \ge 0$,

$$\langle x - x', \partial U(x,y) - \partial U(x',y)\rangle_\gamma = \langle x - x', \nabla U_n(x,y) - \nabla U_n(x',y)\rangle_\gamma + \Delta_n(x,x') \tag{57}$$

where

$$\begin{aligned}|\Delta_n(x,x')| &= |\langle x - x', \partial U(x,y) - \nabla U_n(x,y) + \nabla U_n(x',y) - \partial U(x',y)\rangle_\gamma| \\ &\le \|x-y\|_\gamma\{\|\nabla U_n(x,y) - \partial U(x,y)\|_\gamma + \|\nabla U_n(x',y) - \partial U(x',y)\|_\gamma\}.\end{aligned} \tag{58}$$

Using the facts that $\nabla U_n(\cdot,y) = \nabla U_n(\cdot,y)\circ\Pi_n$ and $\nabla U_n(\cdot,y)$ maps $l_2(\gamma)$ into $l_2^n(\gamma)$, then applying the instance of assumption b) of Proposition 1 which has already been verified for $(\gamma, -\nabla U_n(\cdot,y), n)$ in the proof of Theorem 1, we have for any $x, x' \in l_2(\gamma)$,

$$\begin{aligned}&\langle x - x', \nabla U_n(x,y) - \nabla U_n(x',y)\rangle_\gamma \\ &= \langle \Pi_n(x) - \Pi_n(x'), \nabla U_n(\cdot,y)\circ\Pi_n(x) - \nabla U_n(\cdot,y)\circ\Pi_n(x')\rangle_\gamma \\ &\ge \lambda\|\Pi_n(x) - \Pi_n(x')\|_\gamma^2 \\ &\to \lambda\|x - x'\|_\gamma^2 \quad \text{as} \quad n \to \infty.\end{aligned} \tag{59}$$

By (6), (13) and assumptions a) and b) of Condition 2, we have for any $x \in l_2(\gamma)$,

$$\|\nabla U_n(x,y) - \partial U(x,y)\|_\gamma^2 = \gamma^n \|\nabla_n \tilde{\phi}_n(x,y) - \nabla_n \phi_n(x,y)\|^2 + \sum_{k=n+1}^{\infty} \gamma^k \|\nabla_n \phi_n(x,y)\|^2 \to 0 \quad \text{as} \quad n \to \infty. \tag{60}$$

Combining (57)–(60) gives:

$$\langle x - x', \partial U(x,y) - \partial U(x',y)\rangle_\gamma \geq \lambda \|x - x'\|_\gamma^2, \tag{61}$$

which completes the verification of assumption b) of Proposition 1 for the triple $(\gamma, -\partial U(\cdot, y), \infty)$.

The existence of $\xi_\infty \in l_2(\gamma)$ such that $\partial U(\xi_\infty, y) = 0$, together with (61) implies, via the same arguments as in the proof of Theorem 1, that equations (55) and (56) hold not only for $m \in \mathbb{N}_0$ but also for $m = \infty$. Under assumption a) of Condition 2, the bound: $\eta_n(r,y) \leq \beta_n(y) + \chi r/\gamma$ holds using (8), and plugging in this bound completes the proof of (16).

To see that $\|\xi_n - \xi_\infty\|_\gamma \to 0$ as $n \to \infty$, applying the Cauchy-Schwarz inequality to (11),

$$\|x - x'\|_\gamma^2 \leq \frac{1}{\lambda^2} \|\nabla U_n(x,y) - \nabla U_n(x',y)\|_\gamma^2, \quad \forall x, x' \in l_2^n(\gamma).$$

Choosing $x = \xi_n$ and $x' = \Pi_n(\xi_\infty)$,

$$\begin{aligned}\|\xi_n - \xi_\infty\|_\gamma^2 &= \|\xi_\infty - \Pi_n(\xi_\infty)\|_\gamma^2 + \|\xi_n - \Pi_n(\xi_\infty)\|_\gamma^2 \\ &\leq \|\xi_\infty - \Pi_n(\xi_\infty)\|_\gamma^2 + \frac{1}{\lambda^2} \|\nabla U_n(\Pi_n(\xi_\infty), y)\|_\gamma^2.\end{aligned}$$

Combining with (60), $\xi_\infty \in l_2(\gamma)$ and parts a) and b) of Condition 2 we find that indeed $\|\xi_n - \xi_\infty\|_\gamma \to 0$ as $n \to \infty$.

□

C Proofs for section 4

Proof of Lemma 1. In the setting described in section 4,

$$\begin{aligned}
&\big\langle x_n - x_n', \nabla_n \log f(x_{n-1}, x_n) - \nabla_n \log f(x_{n-1}', x_n')\big\rangle \\
&\quad + \big\langle x_n - x_n', \nabla_n \log f(x_n, x_{n+1}) - \nabla_n \log f(x_n', x_{n+1}')\big\rangle \\
&\quad = -(x_n - x_n')^{\mathrm{T}} (\Sigma^{-1} + A^{\mathrm{T}} \Sigma^{-1} A)(x_n - x_n') + (x_n - x_n')^{\mathrm{T}} \Sigma^{-1} A (x_{n-1} - x_{n-1}') \\
&\quad + (x_n - x_n')^{\mathrm{T}} A^{\mathrm{T}} \Sigma^{-1} (x_{n+1} - x_{n+1}), \\
&\big\langle x_n - x_n', \nabla_n \log f(x_{n-1}, x_n) - \nabla_n \log f(x_{n-1}', x_n')\big\rangle \\
&\quad = -(x_n - x_n')^{\mathrm{T}} \Sigma^{-1} (x_n - x_n') + (x_n - x_n')^{\mathrm{T}} \Sigma^{-1} A (x_{n-1} - x_{n-1}'), \\
&\langle x_0 - x_0', \nabla_0 \log \mu(x_0) - \nabla_0 \log \mu(x_0')\rangle \\
&\quad = -(x_0 - x_0')^{\mathrm{T}} \Sigma_0^{-1} (x_0 - x_0'), \\
&\langle x_0 - x_0', \nabla_0 \log f(x_0, x_1) - \nabla_0 \log f(x_0', x_1')\rangle \\
&\quad = -(x_0 - x_0')^{\mathrm{T}} A^{\mathrm{T}} \Sigma^{-1} A (x_0 - x_0') + (x_0 - x_0')^{\mathrm{T}} A^{\mathrm{T}} \Sigma^{-1} (x_1 - x_1').
\end{aligned}$$

Combining these expressions with (18), (4)-(5) and applying the following bounds:

$$\begin{aligned}\inf_{u \neq 0} \frac{u^{\mathrm{T}} (\Sigma^{-1} + A^{\mathrm{T}} \Sigma^{-1} A) u}{\|u\|^2} &\geq \rho_{\min}(\Sigma^{-1}) + \rho_{\min}(A^{\mathrm{T}} A) \rho_{\min}(\Sigma^{-1}) \\ &= \rho_{\max}(\Sigma)^{-1} \{1 + \rho_{\min}(A^{\mathrm{T}} A)\},\end{aligned} \tag{62}$$

$$\begin{aligned}
&\inf_{u \neq 0} \frac{u^{\mathrm{T}} \Sigma^{-1} u}{\|u\|^2} \wedge \inf_{u \neq 0} \frac{u^{\mathrm{T}} (\Sigma_0^{-1} + A^{\mathrm{T}} \Sigma^{-1} A) u}{\|u\|^2} \\
&\geq \rho_{\min}(\Sigma^{-1}) \wedge \{\rho_{\min}(\Sigma_0^{-1}) + \rho_{\min}(A^{\mathrm{T}} A) \rho_{\min}(\Sigma^{-1})\} \\
&= \rho_{\max}(\Sigma)^{-1} \wedge \{\rho_{\max}(\Sigma_0)^{-1} + \rho_{\min}(A^{\mathrm{T}} A) \rho_{\max}(\Sigma)^{-1}\},
\end{aligned}$$

$$\sup_{u,v\neq 0}\frac{\left|u^{\mathrm{T}}A^{\mathrm{T}}\Sigma^{-1}v\right|}{\|u\|\|v\|}\leq \rho_{\max}(A^{\mathrm{T}}A)^{1/2}\rho_{\max}(\Sigma^{-1})=\rho_{\max}(A^{\mathrm{T}}A)^{1/2}\rho_{\min}(\Sigma)^{-1},$$

gives the expressions for $\zeta,\tilde{\zeta},\theta$ in the statement of the lemma.

□

Proof of Lemma 2. From the Lipschitz assumptions we have:

$$\begin{aligned}
&\|\nabla_n \log f(x_{n-1},x_n)-\nabla_n \log f(x'_{n-1},x'_n)\|\\
&=\|\nabla\psi\{x_n-A(x_{n-1})\}-\nabla\psi\{x'_n-A(x'_{n-1})\}\|\\
&\leq L_{\nabla\psi}\|x_n-x'_n\|+L_{\nabla\psi}L_A\|x_{n-1}-x'_{n-1}\|,\\
&\|\nabla_n \log f(x_n,x_{n+1})-\nabla_n \log f(x'_n,x'_{n+1})\|\\
&\leq \|\nabla A(x_n)\|_{\mathrm{op}}\|\nabla\psi\{x_{n+1}-A(x_n)\}-\nabla\psi\{x'_{n+1}-A(x'_n)\}\|\\
&\quad+\|\nabla\psi\{x'_{n+1}-A(x'_n)\}\|\|\nabla A(x_n)-\nabla A(x'_n)\|_{\mathrm{op}}\\
&\leq L_A L_{\nabla\psi}\|x_{n+1}-x'_{n+1}\|+L_A^2 L_{\nabla\psi}\|x_n-x'_n\|\\
&\quad+L_\psi L_{\nabla A}\|x_n-x'_n\|\\
&=(L_A^2L_{\nabla\psi}+L_\psi L_{\nabla A})\|x_n-x'_n\|+L_{\nabla\psi}L_A\|x_{n+1}-x'_{n+1}\|,\\
&\|\nabla\log\mu(x_0)-\nabla\log\mu(x'_0)\|=\|\nabla\psi_0(x_0)-\nabla\psi_0(x'_0)\|\leq L_{\nabla\psi}\|x_0-x'_0\|.
\end{aligned}$$

The proof is completed by combining these estimates with assumption b) of the Lemma and (4)-(5). □

Proof of Lemma 3.

$$\begin{aligned}
\sum_{n=0}^{\infty}\mathbf{P}\left(\rho^n Z_n\geq 1\right)=\sum_{n=0}^{\infty}\mathbf{P}\left(|Z_n|^s\geq\rho^{-sn}\right)&\leq\sum_{n=0}^{\infty}\rho^{sn}\mathbf{E}\left[|Z_n|^s\right]\\
&\leq\left(\sup_{n\geq 0}\frac{\mathbf{E}[|Z_n|^s]}{(n+1)^p}\right)\sum_{n=0}^{\infty}(n+1)^p\rho^{sn}<\infty,
\end{aligned}\tag{63}$$

where the first inequality is Markov's inequality. The result follows from the Borel-Cantelli lemma. □

Proof of Lemma 4. From (7) and (4)-(5), there exists a finite constant c such that

$$\beta_n(y)\leq c+\|\left.\nabla_x\log g(x,y_n)\right|_{x=0}\|^2.\tag{64}$$

The claim of the lemma that $\sup_{n\geq 0}\rho^n\beta_n(Y)<\infty$ a.s. then follows from the fact that for $a\geq 0$ and $s\in(0,1]$, the function $a\mapsto a^s$ is subadditive, combined with Lemma 3.

Using the same subadditivity again, and (64),

$$\alpha_{\gamma,n}(y)^s\leq\sum_{m=0}^{n}\gamma^{s(n-m)}\beta_m(y)^s\leq\frac{c^s}{1-\gamma^s}+\sum_{m=0}^{n}\gamma^{s(n-m)}\|\left.\nabla_x\log g(x,y_m)\right|_{x=0}\|^{2s}.$$

Combining this bound with the assumption of the lemma and the fact that $\sum_{m=0}^{n}m^p$ grows no faster than n^{p+1} as $n\to\infty$, there exists a finite constant $c'(p)$ such that:

$$\begin{aligned}
\mathbf{E}[\alpha_{\gamma,n}(Y)^s]&\leq\frac{c^s}{1-\gamma^s}+\sum_{m=0}^{n}\mathbf{E}\left[\|\left.\nabla_x\log g(x,Y_m)\right|_{x=0}\|^{2s}\right]\\
&\leq\frac{c^s}{1-\gamma^s}+c'(p)n^{p+1}.
\end{aligned}$$

The claim of the lemma that $\sup_{n\geq 0}\rho^n\alpha_{\gamma,n}(Y)<\infty$ a.s. then follows by applying Lemma 3. The proof is completed via bounds:

$$\rho^n\alpha_{\gamma,n}(Y)=\tilde{\rho}^n\left(\frac{\rho}{\tilde{\rho}}\right)^n\alpha_{\gamma,n}(Y)\leq\tilde{\rho}^n\sup_{m\geq 0}\left(\frac{\rho}{\tilde{\rho}}\right)^m\alpha_{\gamma,m}(Y)$$

and

$$\sum_{k=n}^{\infty}\rho^k\beta_k(Y)=\sum_{k=n}^{\infty}\tilde{\rho}^k\left(\frac{\rho}{\tilde{\rho}}\right)^k\beta_k(Y)\leq\frac{\tilde{\rho}^n}{1-\tilde{\rho}}\sup_{m\geq 0}\left(\frac{\rho}{\tilde{\rho}}\right)^m\beta_m(Y).$$

□

References

[1] A. Caliebe. Properties of the maximum a posteriori path estimator in hidden Markov models. *IEEE transactions on information theory*, 52(1):41–51, 2006.

[2] A. Caliebe and U. Rosler. Convergence of the maximum a posteriori path estimator in hidden Markov models. *IEEE Transactions on Information Theory*, 48(7):1750–1758, 2002.

[3] P. Chigansky and Y. Ritov. On the Viterbi process with continuous state space. *Bernoulli*, 17(2): 609–627, 2011.

[4] K. Deimling. *Ordinary differential equations in Banach spaces*, volume 596 of *Lecture Notes in Mathematics*. Springer, 2006.

[5] C. Donner, K. Obermayer, and H. Shimazaki. Approximate inference for time-varying interactions and macroscopic dynamics of neural populations. *PLoS computational biology*, 13(1):1–27, 2017.

[6] R. Douc and E. Moulines. Asymptotic properties of the maximum likelihood estimation in misspecified hidden Markov models. *The Annals of Statistics*, 40(5):2697–2732, 2012.

[7] R. Douc, G. Fort, E. Moulines, and P. Priouret. Forgetting the initial distribution for hidden Markov models. *Stochastic processes and their applications*, 119(4):1235–1256, 2009.

[8] Randal Douc, Eric Moulines, and David Stoffer. *Nonlinear time series: Theory, methods and applications with R examples*. Chapman and Hall/CRC, 2014.

[9] James Durbin and Siem Jan Koopman. *Time series analysis by state space methods*. Oxford university press, 2012.

[10] C. Hopper and B. Andrews. *The Ricci flow in Riemannian geometry*. Number 2011 in Lecture Notes in Mathematics. Springer, 2011.

[11] James J Jun et al. Fully integrated silicon probes for high-density recording of neural activity. *Nature*, 551(7679):232, 2017.

[12] Alexey Koloydenko and Jüri Lember. Infinite viterbi alignments in the two state hidden markov models. *Acta Comment. Univ. Tartu. Math*, 12:109–124, 2008.

[13] J. Kuljus, K.and Lember. Asymptotic risks of viterbi segmentation. *Stochastic processes and their applications*, 122(9):3312–3341, 2012.

[14] J. Lember and A.A. Koloydenko. A constructive proof of the existence of Viterbi processes. *IEEE Transactions on information theory*, 56(4):2017–2033, 2010.

[15] J.I. Lember and A. Koloydenko. The adjusted Viterbi training for hidden Markov models. *Bernoulli*, 14(1):180–206, 2008.

[16] Jüri Lember and Joonas Sova. Existence of infinite Viterbi path for pairwise Markov models. *Stochastic Processes and their Applications*, 130(3):1388–1425, 2020.

[17] Jüri Lember and Joonas Sova. Regenerativity of Viterbi process for pairwise Markov models. *Journal of Theoretical Probability*, 34(1):1–33, 2021.

[18] A. Saumard and J. A. Wellner. Log-concavity and strong log-concavity: a review. *Statistics surveys*, 8:45–114, 2014.

[19] H. Shimazaki, S. Amari, E. N. Brown, and S. Grün. State-space analysis of time-varying higher-order spike correlation for multiple neural spike train data. *PLoS Computational Biology*, 8:e1002385, 2012.

[20] Mike West and Jeff Harrison. *Bayesian forecasting and dynamic models*. Springer Science & Business Media, 2006.