Recursive Learning of Asymptotic Variational Objectives

Alessandro Mastrototaro
Department of Mathematics
KTH Royal Institute of Technology
Stockholm, Sweden
alemas@kth.se

Mathias Müller
Department of Mathematics
KTH Royal Institute of Technology
Stockholm, Sweden
matmul@kth.se

Jimmy Olsson
Department of Mathematics
KTH Royal Institute of Technology
Stockholm, Sweden
jimmyol@kth.se

Abstract

General *state-space models* (SSMs) are widely used in statistical machine learning and are among the most classical generative models for sequential time-series data. SSMs, comprising latent Markovian states, can be subjected to *variational inference* (VI), but standard VI methods like the *importance-weighted autoencoder* (IWAE) lack functionality for streaming data. To enable online VI in SSMs when the observations are received in real time, we propose maximising an IWAE-type variational lower bound on the asymptotic *contrast function*, rather than the standard IWAE ELBO, using stochastic approximation. Unlike the *recursive maximum likelihood* method, which directly maximises the asymptotic contrast, our approach, called online sequential IWAE (OSIWAE), allows for online learning of both model parameters and a Markovian recognition model for inferring latent states. By approximating filter state posteriors and their derivatives using sequential Monte Carlo (SMC) methods, we create a particle-based framework for online VI in SSMs. This approach is more theoretically well-founded than recently proposed *online variational SMC* methods. We provide rigorous theoretical results on the learning objective and a numerical study demonstrating the method's efficiency in learning model parameters and particle proposal kernels.

1 INTRODUCTION

The *variational autoencoder* (VAE) (Kingma and Welling, 2014) is a foundational probabilistic method in machine learning, renowned for its capability to learn latent-variable models. Building on this framework, the *importance-weighted autoencoder* (IWAE) (Burda et al., 2016) was developed to enhance the performance of probabilistic inference by providing a tighter bound on the marginal likelihood compared to the standard VAE using a multi-sample objective. The VAE and the IWAE have found diverse applications across many fields, including, *e.g.*, deep generative modelling, recommendation systems, biology, and image compression (Zhang et al., 2020; Bond-Taylor et al., 2021; Gayoso et al., 2021; Xu et al., 2022; Daudel et al., 2023; Doucet et al., 2023). In the case where the observed data is modelled by *state-space models* (SSMs), also known as general state-space *hidden Markov models* (HMMs), which is the most classical probabilistic generative modelling framework for time series (see Cappé et al., 2005), *variational sequential Monte Carlo* (VSMC) (Maddison et al., 2017; Le et al., 2018; Naesseth et al., 2018) can be seen as a further development of the IWAE, where the importance-sampling-based estimator of the likelihood used in the IWAE is replaced by the (still unbiased) likelihood estimator provided by *sequential Monte Carlo* (SMC) *methods*, also known as *particle filters* (Doucet et al., 2001; Chopin and Papaspiliopoulos, 2020). VSMC not only enables parameter estimation in SSMs, but also allows for acceleration of SMC by optimising the particle proposal distributions. Such adaptation, which enables reduced particle degeneracy and improved accuracy, is essential to make the SMC method better suited for complex and high-dimensional data-assimilation problems. An inherent, general problem of VSMC is the difficulty to reparameterise the resampling operation of the particle filter and consequently to estimate the corresponding ELBO gradient with acceptable accuracy. This has prompted its creators to use an *ad hoc* truncated version of the gradient, in which terms with high variance are simply excluded, leading to a bias that is difficult to control. Moreover, since the standard implementation of VSMC is primarily designed for batch-processing scenarios, it is not suitable for processing streaming time-series data, which is the focus of this work. An attempt to adapt the VSMC methodology to online scenarios was recently

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

made by Mastrototaro and Olsson (2024), but although this methodology has been shown to work well in some cases, it relies on a time-distributed stochastic gradient derived from its batch-oriented predecessors, and consequently suffers from the same lack of theoretical support.

Thus, in this paper, we take a different approach and focus instead on maximising a lower bound on the *asymptotic contrast function*, alternatively termed the *log-likelihood rate*, given by the ergodic limit of the time-normalised log-likelihood function when the number of observations tends to infinity (see, *e.g.*, Cappé et al., 2005; Tadić and Doucet, 2021), in an online scenario in which the data is only made available sequentially in real time. More specifically, in the proposed algorithm, which we refer to as the *online sequential importance-weighted autoencoder* (OSIWAE), a *contrast lower bound* (COLBO), interpreted as an IWAE-type multi-sample variational objective, is maximised using a Robbins–Monro scheme with Markovian perturbations targeting the zeros of the COLBO gradient. This allows to learn, simultaneously, SSM parameters as well as a Markovian recognition model depending on the given data point and on the previous latent state. The latter can be used as an effective particle proposal or as a sequential encoder similar to the *variational recurrent neural networks* introduced by Chung et al. (2015). In support of OSIWAE, we present theoretical results that establish the COLBO ergodic limit and furthermore provide $\mathcal{O}(M^{-1})$ bounds (where M is the COLBO sample size) on the discrepancies between the COLBO and the asymptotic contrast function as well as between their gradients.

Ideally, OSIWAE requires access to the flow of filter state posteriors and their derivatives, which are however intractable in general. By approximating these measures using particle filters, a practical version of the algorithm, referred to as SMC-OSIWAE, is obtained. In particular, the filter derivative, or *tangent filter*, requires the calculation of expectations of additive state functionals—a problem that has received much attention in the SMC literature (see, *e.g.*, Kitagawa and Sato, 2001; Olsson et al., 2008; Del Moral et al., 2010; Olsson and Westerborn, 2017). Hence, we propose an SMC-based OSIWAE, called SMC-OSIWAE, incorporating the latest advancement in this line of research, namely the *AdaSmooth* algorithm (Mastrototaro et al., 2024), which allows for online approximation of the tangent filters with complexity and memory requirements that only grow linearly with the number of particles. Unlike OVSMC, SMC-OSIWAE has a solid theoretical underpinning. When it comes to learning the model parameters, SMC-OSIWAE approaches the particle-based *recursive maximum likelihood* (RML, Le Gland and Mevel, 1997; Poyiadjis et al., 2011; Del Moral et al., 2015) with increasing M, but where a lower bound on the contrast function is maximised rather than the contrast function itself. On the other hand, when learning is restricted to proposal parameters only, SMC-OSIWAE becomes very similar to OVSMC. In this way, SMC-OSIWAE can be interpreted as a golden compromise between these two methods. Furthermore, the SMC-OSIWAE updating formula sheds some light on the bias of OVSMC (and thus of VSMC), which can be expected to be significant when the observations are noisy relative to the latent state signal. This latter conclusion is also confirmed by our simulations.

Finally, we present numerical experiments to showcase the effectiveness of the algorithm in learning SSM parameters as well as close-to-optimal particle proposals. These experiments highlight OSIWAE's advantages over its predecessors OVSMC and particle-based RML.

The paper is structured as follows. In Section 2 we provide a brief overview of SSMs and introduce the concepts of the asymptotic contrast function and the RML procedure. In Section 3, we introduce the general OSIWAE idea, and illustrate our theoretical results. In Section 4 we describe the particle-based implementation of OSIWAE, SMC-OSIWAE, and in Section 5 we provide some some numerical illustrations of the latter. Details and proofs are found in the appendix.

2 BACKGROUND

2.1 Model and Notation

An SSM is a bivariate Markov chain $(X_t, Y_t)_{t\geq 0}$ evolving on some measurable state space $(\mathsf{X} \times \mathsf{Y}, \mathcal{X} \otimes \mathcal{Y})$, which is typically Euclidean and furnished with the Borel σ-field. Here the so-called *state process* $(X_t)_{t\geq 0}$ is latent, or *hidden*, and only partially observed through the *observation process* $(Y_t)_{t\geq 0}$. For a given time horizon $T \in \mathbb{N}$, the joint law of $(X_t, Y_t)_{t=0}^T$ is $p_\theta(x_{0:T}, y_{0:T}) = m_0(x_0) g_\theta(y_0 \mid x_0) \prod_{t=1}^T m_\theta(x_t \mid x_{t-1}) g_\theta(y_t \mid x_t)$, where m_θ and g_θ are Markov transition densities w.r.t. some reference measures (typically Lebesgue) on $\mathcal{X}$ and $\mathcal{Y}$, respectively, and m_0 is some probability density w.r.t. the same reference measure on $\mathcal{X}$. Here $x_{0:T} = (x_0, \dots, x_T)$ is our generic notation for vectors and $\theta \in \Theta \subset \mathbb{R}^p$, $p \in \mathbb{N}_{>0}$, is a parameter vector. Under this dynamics, the state process is itself Markov with transition density m_θ and initial density m_0. Furthermore, conditionally on $(X_t)_{t=0}^T$, the observations $(Y_t)_{t=0}^T$ are independent with marginals $g_\theta(\cdot \mid X_t)$, $0 \leq t \leq T$. In the practical application of SSMs, the latent states are usually inferred from the observations using the so-called *joint-smoothing distributions* $\phi_{0:t}^\theta(x_{0:t}) := p_\theta(x_{0:t} \mid y_{0:t})$ or *filter distributions* $\phi_t^\theta(x_t) := p_\theta(x_t \mid y_{0:t})$. These state posteriors are also of paramount importance when inferring the parameter θ using maximum likelihood estimation (see, *e.g.*, Cappé et al., 2005, Chapters 10–11).

2.2 The Asymptotic Contrast Function

Given a batch $Y_{0:T}$ of observations, the maximum-likelihood estimator (MLE) of θ is the parameter $\hat{\theta} \in \Theta$

such that $\log p_{\hat{\theta}}(Y_{0:T}) \geq \log p_\theta(Y_{0:T})$ for all $\theta \in \Theta$. However, in this paper we focus on the more challenging situation where the data become available via a data stream $(Y_t)_{t\in\mathbb{N}}$. The data is assumed to be generated by some SSM, which does not necessarily belong to the parametric family governed by θ. In this case, it becomes increasingly costly to evaluate the log-likelihood, up to the point where reprocessing all the observations as soon as a new one is recorded becomes infeasible. We will therefore focus instead on maximising the so-called *contrast function* $\ell : \Theta \ni \theta \mapsto \lim_{t\to\infty} t^{-1} \log p_\theta(Y_{0:t})$ with respect to θ. In order to establish the (a.s.) existence of this objective (see Section 3.2), one typically considers the *extended Markov chain* $(X_{t+1}, Y_{t+1}, \phi_t^\theta, \psi_t^\theta)_{t\in\mathbb{N}}$, where $\psi_t^\theta(x) = \nabla_\theta \phi_t^\theta(x)$ is the so-called *tangent-filter* density. Indeed, the fact that this process is Markov follows from the Markovianity of the SSM and the existence of mappings Φ_θ and Ψ_θ such that for all $t \in \mathbb{N}$, $\phi_{t+1}^\theta = \Phi_\theta(\phi_t^\theta, Y_{t+1})$ and $\psi_{t+1}^\theta = \Psi_\theta(\psi_t^\theta, \phi_t^\theta, Y_{t+1})$ (see Appendix A for details). This chain can be shown to be ergodic under certain mixing assumptions (see, *e.g.*, Le Gland and Mevel, 1997; Douc and Matias, 2001; Tadić and Doucet, 2005), and we denote by Π_θ its stationary distribution and by $\bar{\Pi}_\theta$ the marginal of Π_θ w.r.t. the observation and the filter (it should be remarked that the existence of Π_θ is a mathematically involved topic, and we refer to the previous references for discussions). Using the strong law of large numbers for Markov chains, this construction allows us to express the contrast function as an expectation under Π_θ according to

$$\begin{aligned}\ell(\theta) &= \lim_{t\to\infty} \frac{1}{t} \sum_{s=0}^{t-1} \log \iint g_\theta(Y_{s+1} \mid x_{s+1}) \\ &\qquad \times m_\theta(x_{s+1} \mid x_s)\, dx_{s+1}\, \phi_s^\theta(x_s)\, dx_s \\ &= \lim_{t\to\infty} \frac{1}{t} \sum_{s=0}^{t-1} V_\theta(Y_{s+1}, \phi_s^\theta) = \iint V_\theta(y, \phi)\, \bar{\Pi}_\theta(dy, d\phi)\end{aligned} \tag{1}$$

(a.s.), where we have defined

$$V_\theta(y, \phi) := \log \iint g_\theta(y \mid x') m_\theta(x' \mid x)\, dx'\, \phi(x)\, dx. \tag{2}$$

It has been demonstrated that if the data is assumed to be generated by some SSM in the parametric family of interest, specified by some 'true' parameter $\theta^* \in \Theta$, then, under some identifiability assumptions, $\ell(\theta)$ is maximised by θ^* and the MLE tends tends almost surely to θ^* as t tends to infinity (strong consistency); see, *e.g.*, Douc and Matias (2001, Theorems 1–2) or Cappé et al. (2005, Section 12.4). In `RML`, the contrast function is maximised online using stochastic approximation (Robbins and Monro, 1951). Arguing as in (1), it can be shown that $\nabla_\theta \ell(\theta) = \int G_\theta(y, \phi, \psi)\, \tilde{\Pi}_\theta(dy, d\phi, d\psi)$, where $\tilde{\Pi}_\theta$ is the marginal of Π_θ w.r.t. the observation, the filter, and the tangent filter, and

$$\begin{aligned}G_\theta(y, \phi, \psi) := &\frac{\iint g_\theta(y \mid x') m_\theta(x' \mid x)\, dx' \psi(x)\, dx}{\iint g_\theta(y \mid x') m_\theta(x' \mid x)\, dx' \phi(x)\, dx} \\ &+ \frac{\iint \nabla_\theta \{g_\theta(y \mid x') m_\theta(x' \mid x)\}\, dx' \phi(x)\, dx}{\iint g_\theta(y \mid x') m_\theta(x' \mid x)\, dx' \phi(x)\, dx}.\end{aligned}$$

Now, letting $\nabla_\theta \ell(\theta)$ serve as the mean field of a stochastic approximation scheme with state-dependent Markov noise (see, *e.g.*, Karimi et al., 2019, Case 2), a recursive Robbins–Monro algorithm finding a stationary point of the constrast function is given by

$$\theta_{t+1} \leftarrow \theta_t + \gamma_{t+1} G_{\theta_t}(y_{t+1}, \phi_t^{\theta_{0:t}}, \psi_t^{\theta_{0:t}}), \quad t \in \mathbb{N}, \tag{3}$$

followed by the updates $\phi_{t+1}^{\theta_{0:t+1}} = \Phi_{\theta_{t+1}}(\phi_t^{\theta_{0:t}}, y_{t+1})$ and $\psi_{t+1}^{\theta_{0:t+1}} = \Psi_{\theta_{t+1}}(\psi_t^{\theta_{0:t}}, \phi_t^{\theta_{0:t}}, y_{t+1})$, where $(\gamma_t)_{t\in\mathbb{N}_{>0}}$ is a suitable-chosen sequence of step sizes. The recursion (3) is initialised by some guess θ_0 with associated time-zero filter and tangent filter $\phi_0^{\theta_0}$ and $\psi_0^{\theta_0}$, respectively. Except in cases where the model is linear Gaussian, G_θ, Φ_θ, and Ψ_θ have no closed-form expressions, so the practical implementation of (3) generally requires these quantities to be approximated. For this purpose, SMC methods have proved particularly useful (see, *e.g.*, Poyiadjis et al., 2011; Del Moral et al., 2015; Olsson and Westerborn Alenlöv, 2020), and we shall return to this in Section 4.

3 THE ONLINE SEQUENTIAL IMPORTANCE-WEIGHTED AUTOENCODER (`OSIWAE`)

In the following our goal is to determine a lower bound on the asymptotic contrast by following the principles of the `IWAE` and to design a stochastic-approximation scheme to maximise the same. By maximising a lower bound on the contrast function, we are able to learn not only the model parameters, but also a variational recognition model, which can be used, for example, as a particle proposal. This is not possible through standard `RML`. Similar to `RML`, our first algorithm, `OSIWAE`, outlined in Section 3 will not be implementable in the general case. Therefore, in Section 4 we will present a practical particle-based version, `SMC-OSIWAE`.

3.1 The `OSIWAE` Algorithm

We return to (2) and focus on the inner integral $\int g_\theta(y \mid x') m_\theta(x' \mid x)\, dx'$ of V_θ, which represents the likelihood of a certain observation $y \in \mathsf{Y}$ of the SSM given the latent state $x \in \mathsf{X}$ at the previous time step. Given x and y, we may be interested in inferring the latent state x' at the next time step by determining the conditional distribution $p_\theta(x' \mid x, y) \propto g_\theta(y \mid x') m_\theta(x' \mid x)$. This conditional distribution is of crucial importance in particle filters, since it corresponds to the *locally optimal proposal* (see, *e.g.*, Cornebise et al., 2008, for details). The optimal proposal allows the particles to be guided more efficiently than the

naive *bootstrap proposal* $m_\theta(x' \mid x)$, which mutates the particles 'blindly', without including information about the current observation (Gordon et al., 1993). Alternatively, if the SSM is interpreted as a model for encoding a process on a high-dimensional space Y into a space X of significantly lower dimension, then $p_\theta(x' \mid x, y)$ provides a sequential encoder that takes into account both the observed data and the previously encoded state (see, *e.g.*, Chung et al., 2015). The optimal kernel can be determined in a closed form for only a few model types (Doucet et al., 2000; Cappé et al., 2005, Section 7.2.2.2); thus, our goal is to learn an approximation $r_\theta(x' \mid x, y)$, referred to simply as the *proposal*, of the optimal proposal using variational inference. Note that r_θ is parameterised by the same θ as the SSM. This notation covers both the case where the parameters of the proposal are a subset of the parameters of the SSM as well as the more interesting case where the proposal involves additional parameters. In the latter case, Θ is the product of a model and a proposal-only parameter space.

Now, let r_θ be such that for every $x \in \mathsf{X}$ and $y \in \mathsf{Y}$, $\{x' : r_\theta(x' \mid x, y) = 0\} \subseteq \{x' : m_\theta(x' \mid x) = 0\}$. Moreover, for a given probability density ϕ on X and $y \in \mathsf{Y}$, define the joint probability densities $p_\theta^\phi(x, x', y) := \phi(x) m_\theta(x' \mid x) g_\theta(y \mid x')$ and $q_\theta^\phi(x, x' \mid y) := \phi(x) r_\theta(x' \mid x, y)$. Using this notation and definition (2), write

$$V_\theta(y, \phi) = \log \mathbb{E}_{q_\theta^\phi(\cdot \mid y)} \left[\frac{p_\theta^\phi(X, X', y)}{q_\theta^\phi(X, X' \mid y)} \right].$$

From this expression we immediately see that we are in the framework of `VAE`s (Kingma and Welling, 2014), where p_θ^ϕ and q_θ^ϕ are the *generative* and *recognition models*, respectively, in the special case where the unobserved latent variable comprises two consecutive states. From here it is a short step to generalising the variational objective using the `IWAE` framework (Burda et al., 2016) , yielding

$$\begin{aligned} &V_\theta^M(y, \phi) \\ &\quad := \mathbb{E}_{q_\theta^\phi(\cdot \mid y)^{\otimes M}} \left[\log \left(\frac{1}{M} \sum_{i=1}^M \frac{p_\theta^\phi(X^i, X'^i, y)}{q_\theta^\phi(X^i, X'^i \mid y)} \right) \right] \\ &\qquad\qquad \leq V_\theta(y, \phi), \end{aligned}$$

where $M \in \mathbb{N}_{>0}$ is a sample-size hyperparameter. Here $(X^i, X'^i)_{i=1}^M$ are independent draws from $q_\theta^\phi(\cdot \mid y)$. Thus, we may consider the asymptotic variational objective $\ell^M : \Theta \ni \theta \mapsto \lim_{t\to\infty} t^{-1} \sum_{s=0}^{t-1} V_\theta^M(Y_{s+1}, \phi_s^\theta)$, whose (a.s.) existence is guaranteed by Proposition 3.1 and which, since it bounds the contrast function from below, will be referred to as *contrast lower bound* (COLBO). Note that the COLBO can be expressed as the ergodic limit $\ell^M(\theta) = \int V_\theta^M(y, \phi)\, \bar{\Pi}_\theta(dy, d\phi)$. Now, similarly to the `RML`, we may proceed by constructing a stochastic-approximation scheme with state-dependent Markov noise targeting the zeros of $\nabla_\theta \ell^M(\theta)$, which in turn coincides with the (a.s.) limit of $t^{-1} \sum_{s=0}^{t-1} \nabla_\theta V_\theta^M(Y_{s+1}, \phi_s^\theta)$; see again Proposition 3.1. In order to identify the stochastic update, we need to derive an explicit expression for $\nabla_\theta V_\theta^M(Y_{t+1}, \phi_t^\theta)$ as a function of the states of the extended Markov chain. In this derivation, we will apply the reparameterisation trick (Kingma and Welling, 2014), by assuming that there exists some auxiliary random variable U, taking on values in some measurable space $(\mathsf{U}, \mathcal{U})$ and having distribution $\nu(u)$ on $(\mathsf{U}, \mathcal{U})$ (the latter not depending on θ), and some function h_θ on $\mathsf{X} \times \mathsf{Y} \times \mathsf{U}$, parameterised by θ and differentiable with respect to the same for any given argument (x, y, u), such that for every $(x, y) \in \mathsf{X} \times \mathsf{Y}$, the pushforward distribution $\nu \circ h_\theta^{-1}(x, y, \cdot)$ coincides with that governed by $r_\theta(\cdot \mid x, y)$. Defining the weight function

$$w_\theta(x, y, u) := \frac{g_\theta(y \mid h_\theta(x, y, u)) m_\theta(h_\theta(x, y, u) \mid x)}{r_\theta(h_\theta(x, y, u) \mid x, y)}$$

allows us to write, for a given $y \in \mathsf{Y}$,

$$\begin{aligned} \nabla_\theta V_\theta^M(y, \phi_t^\theta) &= \nabla_\theta \iint \log \left(\frac{1}{M} \sum_{i=1}^M w_\theta(x^i, y, u^i) \right) \\ &\qquad \times \prod_{j=1}^M \nu(u^j) \phi_t^\theta(x^j)\, du^{1:M}\, dx^{1:M} \\ &= \mathbb{E}_{(\phi_t^\theta \otimes \nu)^{\otimes M}} \left[\frac{\sum_{i=1}^M \nabla_\theta w_\theta(X^i, y, U^i)}{\sum_{i'=1}^M w_\theta(X^{i'}, y, U^{i'})} \right] \\ &\quad + \sum_{j'=1}^M \iint \log \left(\frac{1}{M} \sum_{i=1}^M w_\theta(x^i, y, u^i) \right) \nabla_\theta \phi_t^\theta(x^{j'}) \\ &\qquad \times \prod_{\substack{j=1 \\ j \neq j'}}^M \phi_t^\theta(x^j) \prod_{k=1}^M \nu(u^k)\, du^{1:M}\, dx^{1:M}, \end{aligned}$$

where, in the first term, $(X^i, U^i)_{i=1}^M$ are i.i.d. with distribution $\phi_t^\theta(x)\nu(u)$. Note that by symmetry, the terms of the outer sum are identical; hence, letting

$$\begin{aligned} G_\theta^M(y, \phi, \psi) &:= \mathbb{E}_{(\phi \otimes \nu)^{\otimes M}} \left[\frac{\sum_{i=1}^M \nabla_\theta w_\theta(X^i, y, U^i)}{\sum_{i'=1}^M w_\theta(X^{i'}, y, U^{i'})} \right] \\ &\quad + M \int \mathbb{E}_{(\phi \otimes \nu)^{\otimes (M-1)}} \left[\log \left(\frac{1}{M} w_\theta(x, y, u) \right. \right. \\ &\qquad \left. \left. + \frac{1}{M} \sum_{i=1}^{M-1} w_\theta(X^i, y, U^i) \right) \right] \psi(x) \nu(u)\, du\, dx, \end{aligned}$$

we may write $\nabla_\theta V_\theta^M(y_{t+1}, \phi_t^\theta) = G_\theta^M(y_{t+1}, \phi_t^\theta, \psi_t^\theta)$ and, consequently, $\nabla_\theta \ell^M(\theta) = \int G_\theta^M(y, \phi, \psi)\, \bar{\Pi}_\theta(dy, d\phi, d\psi)$. Thus, similar to (3), we may find a zero of $\nabla_\theta \ell^M(\theta)$ using the Robbins–Monro scheme

$$\theta_{t+1} \leftarrow \theta_t + \gamma_{t+1} G_{\theta_t}^M(Y_{t+1}, \phi_t^{\theta_{0:t}}, \psi_t^{\theta_{0:t}}), \quad t \in \mathbb{N}, \quad (4)$$

followed by the updates $\phi_{t+1}^{\theta_{0:t+1}} = \Phi_{\theta_{t+1}}(\phi_t^{\theta_{0:t}}, Y_{t+1})$ and $\psi_{t+1}^{\theta_{0:t+1}} = \Psi_{\theta_{t+1}}(\psi_t^{\theta_{0:t}}, \phi_t^{\theta_{0:t}}, Y_{t+1})$, where $(\gamma_t)_{t \in \mathbb{N}_{>0}}$ is a

suitable-chosen sequence of step sizes. We refer to schedule (4), which is initialised as the RML (3), as the *online sequential importance-weighted auto-encoder* (OSIWAE). As in the case of RML, a practical implementation requires the approximations of G_θ^M, Φ_θ, and Ψ_θ. This is the objective of Section 4, where we describe a practical version of the OSIWAE based on SMC methods.

3.2 Theoretical Properties of the COLBO

All the results displayed below are established under strong mixing assumptions on the SSM and the data-generating process. Furthermore, m_θ, g_θ, r_θ, and their compositions with the reparameterisation function h_θ are assumed be differentiable in θ with bounded gradients. These assumptions are standard in the literature and point to applications where the state and parameter spaces are compact. All details and proofs are found in Appendix A. Our first result serves to define properly the objectives under consideration.

Proposition 3.1. *For all $M \in \mathbb{N}_{>0}$ there exist real-valued differentiable functions ℓ and ℓ^M on Θ such that for all $\theta \in \Theta$, $\mathbb{P}$-a.s.,*

$$\lim_{t\to\infty} \frac{1}{t} \sum_{s=0}^{t-1} V_\theta(Y_{s+1}, \phi_s^\theta) = \ell(\theta),$$
$$\lim_{t\to\infty} \frac{1}{t} \sum_{s=0}^{t-1} G_\theta(Y_{s+1}, \phi_s^\theta, \psi_s^\theta) = \nabla_\theta \ell(\theta),$$
$$\lim_{t\to\infty} \frac{1}{t} \sum_{s=0}^{t-1} V_\theta^M(Y_{s+1}, \phi_s^\theta) = \ell^M(\theta),$$
$$\lim_{t\to\infty} \frac{1}{t} \sum_{s=0}^{t-1} G_\theta^M(Y_{s+1}, \phi_s^\theta, \psi_s^\theta) = \nabla_\theta \ell^M(\theta).$$

The next result characterizes the OSIWAE objective by establishing the relation between the asymptotic contrast function and the COLBO. It is a direct consequence of (Burda et al., 2016, Theorem 1) and (Nowozin, 2018, Proposition 1).

Proposition 3.2. *For all $\theta \in \Theta$ and $M \in \mathbb{N}_{>0}$, $\ell(\theta) \geq \ell^{M+1}(\theta) \geq \ell^M(\theta)$. Moreover, $\ell(\theta) - \ell^M(\theta)$ is $\mathcal{O}(M^{-1})$ uniformly in θ.*

Finally, we establish an $\mathcal{O}(M^{-1})$ bias between the stochastic gradients G_θ and G_θ^M.

Theorem 3.3. *For all $(y_t)_{t\in\mathbb{N}}$, $G_\theta^M(y_{t+1}, \phi_t^\theta, \psi_t^\theta) - G_\theta(y_{t+1}, \phi_t^\theta, \psi_t^\theta)$ is $\mathcal{O}(M^{-1})$ uniformly in t and θ. In addition, $\nabla_\theta \ell^M(\theta) - \nabla_\theta \ell(\theta)$ is $\mathcal{O}(M^{-1})$ uniformly in θ.*

4 SMC-BASED OSIWAE (SMC−OSIWAE)

We now present an implementable version of the OSIWAE algorithm based on SMC methods. For this purpose, we first provide an alternative expression of G_θ^M, where the integral involving ψ_t^θ is expressed as an expectation of the *complete-data score* $\nabla_\theta \log p_\theta(X_{0:t}, Y_{0:t})$ under the joint-smoothing distribution $\phi_{0:t}^\theta$. The complete-data score is of additive form, allowing for sequential updates with constant complexity (see next section). The following lemma, whose proof is found in Appendix B, summarises these properties. First, for $t \in \mathbb{N}_{>0}$ and $\theta \in \Theta$, define the functions $\varphi_0^\theta(x_0) = \nabla_\theta \log g_\theta(y_0 \mid x_0)$ and

$$\varphi_t^\theta(x_t) := \int \nabla_\theta \log p_\theta(x_{0:t}, y_{0:t}) \, p_\theta(x_{0:t-1} \mid y_{0:t-1}, x_t)\, dx_{0:t-1}.$$

Moreover, let, for every density ϕ, function φ, and $y \in \mathsf{Y}$,

$$\bar{G}_\theta^M(y, \phi, \varphi) := \mathbb{E}_{(\phi\otimes\nu)^{\otimes M}} \left[\frac{\sum_{i=1}^M \nabla_\theta w_\theta(X^i, y, U^i)}{\sum_{i'=1}^M w_\theta(X^{i'}, y, U^{i'})} + M \log\left(\frac{1}{M} \sum_{i=1}^M w_\theta(X^i, y, U^i) \right) \times \left(\varphi(X^M) - \mathbb{E}_\phi[\varphi(X)] \right) \right]. \quad (5)$$

Then the following holds true.

Lemma 4.1.

(i) There exists a mapping $\bar{\Psi}_\theta$ such that for all $t \in \mathbb{N}$, $\varphi_{t+1}^\theta = \bar{\Psi}_\theta(\varphi_t^\theta, \phi_t^\theta, Y_{t+1})$.

(ii) It holds that $\bar{G}_\theta^M(Y_{t+1}, \phi_t^\theta, \varphi_t^\theta) = G_\theta^M(Y_{t+1}, \phi_t^\theta, \psi_t^\theta)$.

Building on Lemma 4.1, the OSIWAE procedure may be reformulated by substituting G_θ^M with $\bar{G}_\theta^M$ and replacing the tangent-filter sequence by $(\varphi_t^{\theta_{0:t}})_{t\in\mathbb{N}}$. These updates are performed online as well according to $\varphi_{t+1}^{\theta_{0:t+1}} = \bar{\Psi}_{\theta_{t+1}}(\varphi_t^{\theta_{0:t}}, \phi_t^{\theta_{0:t}}, Y_{t+1})$. The initialisation step involves computing $\phi_0^{\theta_0}$ and setting $\varphi_0^{\theta_0}(x_0) = \nabla_\theta \log g_{\theta_0}(Y_0 \mid x_0)$. Still, this idealised approach is impractical for direct implementation, why a particle-based version of the same is presented in the next section.

4.1 OSIWAE Gradient-Step Approximation

We assume that at each time $t \in \mathbb{N}$ we have access to some weighted particle sample $(\xi_t^i, \omega_t^i)_{i=1}^N$, $N \in \mathbb{N}_{>0}$, whose associated weighted empirical measure approximates ϕ_t^θ. In addition, assume that we have access to some associated statistics $(\tau_t^i)_{i=1}^N$ such that $\tau_t^i \simeq \varphi_t^\theta(\xi_t^i)$ for all i. We assume that $\sum_{i=1}^N \omega_t^i f(\xi_t^i)/\Omega_t \simeq \mathbb{E}_{\phi_t^\theta}[f(X)]$ and $\sum_{i=1}^N \omega_t^i \tau_t^i f(\xi_t^i)/\Omega_t \simeq \mathbb{E}_{\phi_t^\theta}[f(X)\varphi_t^\theta(X)]$, where $\Omega_t := \sum_{i=1}^N \omega_t^i$, for every measurable function f such that these expectations are well defined. The sample $(\xi_t^i, \tau_t^i, \omega_t^i)_{i=1}^N$ will be produced using the so-called AdaSmooth algorithm proposed by Mastrototaro et al. (2024) (see Appendix C).

Given this sample, our goal is to approximate the expectation (5) when the inputs ϕ_t^θ and φ_t^θ are replaced by their particle approximations. In the standard IWAE, when M is sufficiently large, a good estimate of the gradient is typically obtained by simply drawing M i.i.d. samples from the (reparameterised) recognition model. Thus, in our case we would ideally need M i.i.d. samples from $\phi_t^\theta \otimes \nu$ at the iteration t. However, since ϕ_t^θ is generally intractable, we sample instead M conditionally i.i.d. draws from the empirical distribution formed by a particle sample $(\xi_t^i, \omega_t^i)_{i=1}^N$ targeting ϕ_t^θ. More precisely, write

$$\begin{aligned}
&\bar{G}_\theta^M(y, \phi_t^\theta, \varphi_t^\theta) \\
&= \iint \mathbb{E}_{(\phi_t^\theta \otimes \nu)^{\otimes(M-1)}} \left[\Gamma_1^\theta(x, u, X^{1:M-1}, U^{1:M-1}, y)\right] \\
&\quad + \mathbb{E}_{(\phi_t^\theta \otimes \nu)^{\otimes(M-1)}} \left[\Gamma_2^\theta(x, u, X^{1:M-1}, U^{1:M-1}, y)\right] \\
&\qquad \times (\varphi_t^\theta(x) - \mathbb{E}_{\phi_t^\theta}[\varphi_t^\theta(X)])\, \phi_t^\theta(x) \nu(u)\, dx\, du, \quad (6)
\end{aligned}$$

where

$$\begin{aligned}
&\Gamma_1^\theta(x, u, x^{1:M-1}, u^{1:M-1}, y) \\
&:= \frac{\nabla_\theta w_\theta(x, y, u) + \sum_{i=1}^{M-1} \nabla_\theta w_\theta(x^i, y, u^i)}{w_\theta(x, y, u) + \sum_{k=1}^{M-1} w_\theta(x^k, y, u^k)}, \\
&\Gamma_2^\theta(x, u, x^{1:M-1}, u^{1:M-1}, y) \\
&:= M \log\left(\frac{1}{M} w_\theta(x, y, u) + \frac{1}{M} \sum_{i=1}^{M-1} w_\theta(x^i, y, u^i)\right).
\end{aligned}$$

Now, estimating (i) the inner expectations of (6) based on $M-1$ independent draws $(\check{\xi}_t^i, \check{\upsilon}_t^i)_{i=1}^{M-1}$ generated as

$$(\check{\xi}_t^i, \check{\upsilon}_t^i) \sim \left(\sum_{i=1}^N \frac{\omega_t^i}{\Omega_t} \delta_{\xi_t^i}\right) \otimes \nu,$$

i.e., by resampling pairs of particles and associated statistics in proportion to their weights and providing each resampled pair with a draw from ν, and then (ii) the outer integral using samples $(\hat{\xi}_t^i, \hat{\tau}_t^i, \hat{\upsilon}_t^i)_{i=1}^N$ drawn independently according to

$$(\hat{\xi}_t^i, \hat{\tau}_t^i, \hat{\upsilon}_t^i) \sim \left(\sum_{i=1}^N \frac{\omega_t^i}{\Omega_t} \delta_{(\xi_t^i, \tau_t^i)}\right) \otimes \nu,$$

allows $\bar{G}_\theta^M(y, \phi_t^\theta, \varphi_t^\theta)$ to be estimated by

$$\begin{aligned}
&\bar{\Gamma}^\theta(\hat{\xi}_t^{1:N}, \hat{\upsilon}_t^{1:N}, \check{\xi}_t^{1:M-1}, \check{\upsilon}_t^{1:M-1}, y) \\
&\quad := \frac{1}{N} \sum_{i=1}^N \left\{ \Gamma_1^\theta(\hat{\xi}_t^j, \hat{\upsilon}_t^j, \check{\xi}_t^{1:M-1}, \check{\upsilon}_t^{1:M-1}, y) \right. \\
&\left. + \Gamma_2^\theta(\hat{\xi}_t^j, \hat{\upsilon}_t^j, \check{\xi}_t^{1:M-1}, \check{\upsilon}_t^{1:M-1}, y) \left(\hat{\tau}_t^i - \frac{1}{N} \sum_{\ell=1}^N \hat{\tau}_t^\ell\right) \right\}. \quad (7)
\end{aligned}$$

The procedure describing the practical OSIWAE, referred to as SMC-OSIWAE, is displayed in Algorithm 1, which also includes the online update of $(\xi_t^i, \tau_t^i, \omega_t^i)_{i=1}^N$ via the AdaSmooth online particle smoother described in detail in Appendix C. As we mentioned earlier, in typical applications, $\theta = (\theta^{(1)}, \theta^{(2)})$, where $\theta^{(1)}$ parameterises the SSM only, while $\theta^{(2)}$ parameterises the proposal. Now, note that by Theorem 3.3, G_θ^M converges to G_θ as M tends to infinty, where the latter does not involve $r_{\theta^{(2)}}$; hence, the components of G_θ^M corresponding to the gradient with respect to $\theta^{(2)}$ converge to zero. This becomes a problem when implementing the SMC-OSIWAE, as the estimate of the gradient with respect to $\theta^{(2)}$ suffers from a low signal-to-noise ratio when M is moderately large (we refer to Rainforth et al., 2018, for a discussion), while it is always favourable to use a large M in the model-parameter estimation. Thus, in practice we suggest to repeat twice Algorithm 1 (except for Line 4, which is executed only once) at each iteration t: first with M small, typically equal to 5 or 10, and updating $\theta^{(2)}$ only, then with M large to update $\theta^{(1)}$. Alternative solutions have been discussed by Roeder et al. (2017); Tucker et al. (2018); Finke and Thiery (2019). It is interesting to note that when dealing with $\theta^{(2)}$, since $\nabla_{\theta^{(2)}} \log p_{\theta^{(1)}}(x_{0:t}, y_{0:t}) = 0$, $(\tau_t^i)_{i=1}^N$ are all zero, and so is the second term of (7). In this case, the update of $\theta^{(2)}$ is similar to that performed by the OVSMC method (Mastrototaro and Olsson, 2024, Algorithm 2). However, OVSMC updates the model parameters without the complete-data score term, which is a source of bias of OVSMC. Therefore, although the two methods are derived from different starting points, they can be related. Still, OVSMC lacks a clear asymptotic objective, relying on a hard-to-control truncation of its gradient, whereas OSIWAE aims to maximise a well-defined lower bound on the contrast function, at the price of having to update recursively the statistics $(\tau_t^i)_{i=1}^N$ (as discussed in Appendix C).

Algorithm 1 SMC-OSIWAE

Require: $(\xi_t^i, \tau_t^i, \omega_t^i)_{i=1}^N$, Y_{t+1}, θ_t, step size γ_{t+1}.

1: draw $(\check{\xi}_t^i, \check{\upsilon}_t^i)_{i=1}^{M-1} \overset{\text{i.i.d}}{\sim} \left(\sum_{i=1}^N \frac{\omega_t^i}{\Omega_t} \delta_{\xi_t^i}\right) \otimes \nu$

2: draw $(\hat{\xi}_t^i, \hat{\tau}_t^i, \hat{\upsilon}_t^i)_{i=1}^N \overset{\text{i.i.d}}{\sim} \left(\sum_{i=1}^N \frac{\omega_t^i}{\Omega_t} \delta_{(\xi_t^i, \tau_t^i)}\right) \otimes \nu$

3: set $\theta_{t+1} \leftarrow \theta_t$
$\quad + \gamma_{t+1} \bar{\Gamma}^{\theta_t}(\hat{\xi}_t^{1:N}, \hat{\upsilon}_t^{1:N}, \check{\xi}_t^{1:M-1}, \check{\upsilon}_t^{1:M-1}, Y_{t+1})$

4: run $(\xi_{t+1}^i, \tau_{t+1}^i, \omega_{t+1}^i)_{i=1}^N$
$\quad \leftarrow$ AdaSmooth$((\xi_t^i, \tau_t^i, \omega_t^i)_{i=1}^N, Y_{t+1}, \theta_{t+1})$

5: **return** $(\xi_{t+1}^i, \tau_{t+1}^i, \omega_{t+1}^i)_{i=1}^N$, θ_{t+1}

A rigorous theoretical study of SMC-OSIWAE remains an open challenge, as it requires establishing a bound on the bias of the AdaSmooth-based tangent-filter estimator for a finite number of particles—a nontrivial task. In Tadić and Doucet (2021), a time-uniform $\mathcal{O}(1/N)$ bias is proven for a particle-based tangent-filter estimator derived from the online smoothing algorithm of Del Moral et al. (2010). We are confident that a similar analysis can be conducted for AdaSmooth, supporting our expectation that the bias

of $\bar{\Gamma}^\theta$ with respect to $\bar{G}_\theta^M$ is also time-uniform and of order $\mathcal{O}(1/N)$. Furthermore, Theorem 3.3 states that the bias of $\bar{G}_\theta^M$ with respect to G_θ is $\mathcal{O}(1/M)$. Consequently, we expect the total bias of $\bar{\Gamma}^\theta$ with respect to G_θ to be $\mathcal{O}(1/M + 1/N)$, uniformly in time and θ. This suggests that choosing $N = M$ balances the biases introduced by the importance-weighted estimator in OSIWAE and the particle-based estimator in the SMC scheme. In practice, our experiments indicate that choosing M different from N provides no significant advantage in learning the model parameters.

5 NUMERICAL EXPERIMENTS

In this section, we provide numerical simulations to illustrate the performance of the proposed OSIWAE algorithm in the contexts of parameter learning, optimal filtering, and proposal adaptation. All the experiments were performed on a MacBook Air M2 and used the ADAM optimiser (Kingma and Ba, 2015). If not otherwise stated, a constant learning rate of $\gamma_t = 0.001$ was used.

5.1 Multivariate Linear Gaussian SSM

We consider a 10-dimensional multivariate linear Gaussian SSM to provide an initial assessment of the performance of SMC−OSIWAE. Formally, let the state and observation spaces be $\mathsf{X} = \mathbb{R}^{d_x}$ and $\mathsf{Y} = \mathbb{R}^{d_y}$, respectively, with $d_x = d_y = 10$. The SSM is defined by the state transition density $m_\theta(x_{t+1} \mid x_t) = N_{d_x}(x_{t+1}; Ax_t, S_u S_u^\intercal)$ and the observation density $g_\theta(y_t \mid x_t) = N_{d_y}(y_t; Bx_t, S_v S_v^\intercal)$. Here $A \in \mathbb{R}^{d_x \times d_x}$ and $B \in \mathbb{R}^{d_y \times d_x}$ are the state transition and observation matrices, respectively. The matrices S_u and S_v are diagonal covariance matrices for the process and observation noises. We employ a Gaussian proposal distribution $r_\theta(\cdot|x_t, y_{t+1})$ with mean vector and diagonal covariance matrix parameterised by two distinct neural networks taking x_t and y_{t+1} as inputs. We let the true matrices A and B be diagonal with entries sampled uniformly from $[0.5, 1]$ and generate a dataset of observations by simulating the SSM. We then apply the SMC−OSIWAE algorithm to estimate these matrices while computing particle-based filter expectations of the latent states. We compare the performance of SMC−OSIWAE with the AdaSmooth-based RML method and the OVSMC algorithm. Reference values for the optimal filtering were obtained by executing the Kalman filter for the true model dynamics.

In Figure 1, we see that the parameter estimates produced using SMC−OSIWAE converge faster and exhibit lower MAE compared to OVSMC. This is explained by the fact that the stochastic gradient of OSIWAE incorporates information from the past by estimating the complete-data score, which improves accuracy, especially when the observations are non-informative. It should be noticed that SMC−OSIWAE is almost on par with RML, although SMC−OSIWAE simultaneously adapts the proposal while learning the model parameters. Next, we compare our SMC−OSIWAE approach

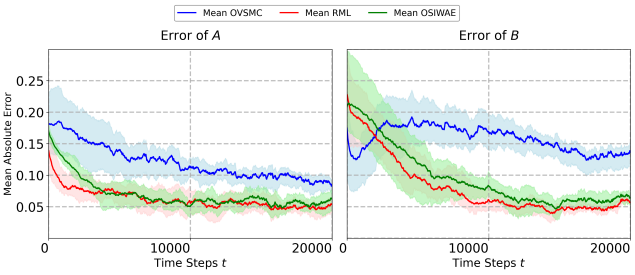

Figure 1: Parameter estimation errors over time for SMC−OSIWAE, OVSMC, and RML in the scenario where $S_u = 0.2I$ and $S_v = 0.5I$. SMC−OSIWAE and RML used $N = 1000$ particles and $M = 1000$ importance samples, while OVSMC used $N = 10000$ particles to ensure comparable computational complexity. The proposal distribution r_θ (a 10-dimensional Gaussian distribution) was parameterised by two single-layer neural networks with 64 nodes each and ReLU activations and learned using $L = 5$ particles. The error bounds are based on 30 independent runs of each algorithm. With our implementation, SMC−OSIWAE took on average 42 min, OVSMC 26 min, and RML 46 min.

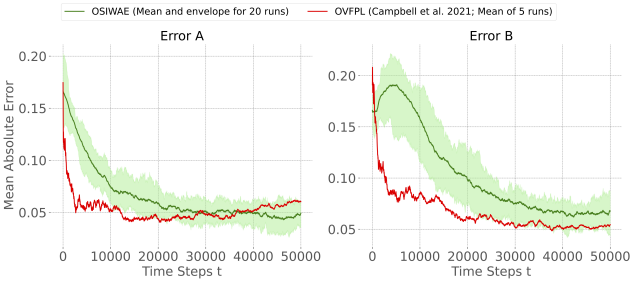

Figure 2: Mean absolute errors of SMC−OSIWAE (green) and (OVFPL) (red). The shaded regions represent the uncertainty envelope for SMC−OSIWAE based on 20 runs, while (OVFPL) shows the mean of 5 runs. SMC−OSIWAE was evaluated using $N = M = 10000$ particles and with the same noise matrices $S_u = 0.2I$ and $S_v = 1.2I$. While (OVFPL) exhibits faster initial convergence, SMC−OSIWAE achieves comparable error over time while being computationally more efficient.

to the *online variational filtering and parameter learning* (OVFPL) method introduced by Campbell et al. (2021). Both methods were evaluated on the same multivariate linear Gaussian SSM. As shown in Figure 2, SMC−OSIWAE achieves parameter estimates with accuracy comparable to that of OVFPL. While OVFPL exhibits faster initial convergence, it relies on a non-amortized variational family, requiring the solution of a regression problem at every time step. This added flexibility comes at the cost of significantly higher computational expense. In contrast, SMC−OSIWAE learns a single, general proposal for all time steps, leading

to a more efficient implementation. As a result, while the public implementation of OVFPL required approximately 17–18 hours per run, our approach only took about 9 hours per run. These findings suggest that SMC-OSIWAE is a strong competitor to OVFPL, particularly given its simplicity of implementation and lower computational cost.

Hence, we next examine the filter-mean estimates produced by these algorithms as the parameters are being learned. We evaluate the MSEs of the filter-mean estimates of each algorithm with respect to the output of the Kalman filter executed for the true model parameters and display the result in Figure 3. Clearly, after an initial phase, when both SMC-OSIWAE and OVSMC learn the proposal parameters and therefore perform worse than RML, SMC-OSIWAE shows a significantly better performance than its competitors in the long run.

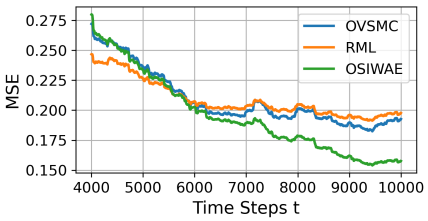

Figure 3: MSEs over time for OSIWAE, OVSMC, and RML with respect to the Kalman filter (executed for true parameters) for the linear Gaussian model with $S_u = 0.5I$ and $S_v = 0.2I$. The values are plotted as moving averages with a window of 3000 time steps. For all methods, the MSEs are based on 50 independent runs on the same data and different starting values of A and B.

5.2 Simultaneous Localisation and Mapping

The *simultaneous localisation and mapping* (SLAM) problem is fundamental in robotics and requires online inference; see, *e.g.*, Dissanayake et al. (2001); Thrun et al. (2005). The goal is to jointly estimate the trajectory of a robot and the positions of $L \in \mathbb{N}_{>0}$ unknown landmarks based on noisy observations. In this context, the latent states are the positions of the robot in a two-dimensional landscape at different time steps. These positions are partially observed through a vector of pairs indicating the distance and the angle with respect to the landmarks. We let $\theta = (\theta^1, \dots, \theta^L)$ be the positions of the landmarks, where $\theta^i = (\theta^i_1, \theta^i_2) \in \mathbb{R}^2$. The robot's motion is modeled as a bivariate random walk with covariance matrix $\sigma^2_{\text{motion}} I_2$. Here we have $Y_t = (Y_t^1, \dots, Y_t^L)$, where Y_t^i is a tuple indicating a noisy measurement of the distance and the angle of the robot with respect to landmark i, for $i \in \{1, \dots L\}$. More specifically, $Y_t^i = h(X_t, \theta^i) + \sigma_{\text{obs}} V_t^i$, where $(V_t^1)_{t\in\mathbb{N}}, \dots, (V_t^L)_{t\in\mathbb{N}}$ are sequences of bivariate i.i.d standard Gaussian random variables. The measurement function is such that $h(x, \theta^i) = (\|\theta^i - x\|,\ \text{atan2}(\theta^i_2 - x_2,\ \theta^i_1 - x_1))$ for all $x = (x_1, x_2) \in \mathbb{R}^2$.

In this experiment, we aim to learn the unknown positions of the landmarks while sequentially estimating the position of the robot. We assume that the noise parameters are known. Figure 4 shows that after an initial phase where SMC-OSIWAE is adapting the proposal, the landmark estimation becomes clearly better compared to both RML and OVSMC, and our algorithm is able to estimate the exact locations more precisely. On the right, we see that SMC-OSIWAE is also able to first train the proposal in some environment and, when used in another one, the learning curve is more accurate than RML and OVSMC from the beginning.

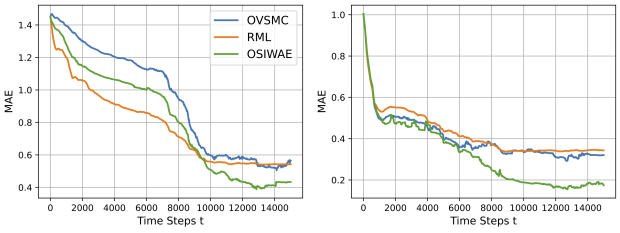

Figure 4: Average MAE of the estimated positions of $L = 8$ landmarks over time using OSIWAE, RML, and OVSMC in a SLAM scenario with $\sigma^2_{\text{motion}} = 0.2$ and $\sigma^2_{\text{obs}} = 0.1$. The proposal distribution $r_\theta(\cdot \mid x_t, y_{t+1})$ in both OSIWAE and OVSMC is learned via two distinct neural networks, each with one hidden layer of 128 nodes. All three methods use $N = 1000$ particles and OSIWAE uses $M = 1000$. Left panel: All three algorithms run on the same data, without any prior learning. Right panel: A training run is first performed using SMC-OSIWAE on a different data record to learn the proposal distribution; afterwards, all three algorithms are applied to the same data.

5.3 Growth Model

Finally, we consider the so-called *growth model* Kitagawa (1987), which is a standard benchmark model in particle filtering due to the highly nonlinear latent process. The state dynamics is given by $X_t = a_{t-1}(X_{t-1}) + \sigma_u U_{t-1}$, where $a_{t-1}(x) = \alpha_0 x + \alpha_1 x/(1+x^2) + \alpha_2 \cos(1.2(t-1))$, and the observations process satisfies $Y_t = bX_t^2 + \sigma_v V_t$, where $(U_t)_{t\in\mathbb{N}}$ and $(V_t)_{t\in\mathbb{N}}$ are i.i.d standard Gaussian random variables. Here, we first generated data with $\alpha_0 = 0.5$, $\alpha_1 = 25$, $\alpha_2 = 8$, $\sigma_u^2 = 10$, $b = 0.05$, and $\sigma_v^2 = 1$; then we used SMC-OSIWAE to estimate α_0, b, and σ_u and simultaneously adapted the particle filter proposal. The interesting aspect of this model is that under certain parameterisations—like the one given—the locally optimal proposal is bimodal, with one dominating mode and the other one almost negligible. In these scenarios, the bootstrap proposal tends to be too diffuse, resulting in many wasted samples. Thus, we run the SMC-OSIWAE algorithm to estimate the unknown parameters of the model while simultaneously learning a

better proposal distribution. We design a family of proposals that integrate new parameters with the ones of the model. This is done by letting again $r_\theta(\cdot \mid x_t, y_{t+1})$ be Gaussian with mean and variance parameterised by neural networks; however, at each time $t \in \mathbb{N}$, in addition to the new observation y_{t+1}, we input the mean $a_t(x_t)$ instead of the current state x_t.

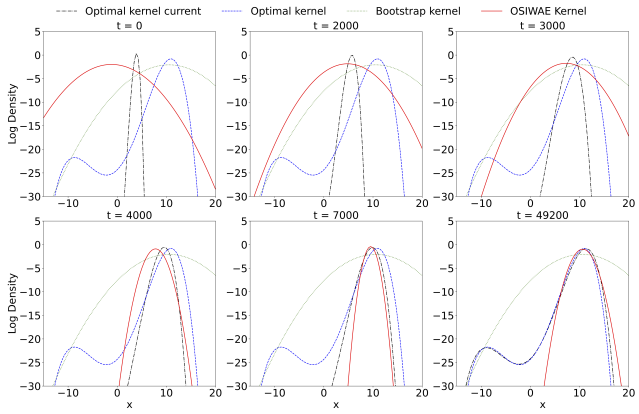

Figure 5: Log-densities of the learned proposal, the optimal kernel parameterised by the current parameter fit as well as the true parameters, and the prior kernel with true parameters. SMC−OSIWAE uses 1000 particles and $M = 1000$. The Gaussian proposal r_θ is parameterised by two distinct neural networks, with one hidden layer of 12 nodes each, modelling the mean and the variance of the same. In each plot, $x_t = 0.1$ and $y_{t+1} = 6$.

Figure 5 illustrates the progression of the proposal distribution. Initially, after a few thousand iterations, the learned proposal starts to approximate the locally optimal kernel, despite the model parameters not yet being fully learned. As the optimal kernel converges to reflect the true parameters, our Gaussian proposal accurately matches the dominant mode. In contrast, the prior kernel of the standard bootstrap particle filter remains overly dispersed.

Figure 5 illustrates a timestep where the growth-model optimal kernel exhibits a single dominant mode. However, under a narrow range of parameter settings, the optimal kernel can display a bimodal distribution of equal magnitude. A rough estimate suggests that this occurs in approximately 5% of the time steps. Although our method employs a univariate proposal distribution, Figure 6 demonstrates that SMC−OSIWAE adapts by broadening the proposal when faced with a bimodal scenario. In such cases, the mean of the learned proposal is positioned between the two modes of the optimal kernel, resulting in a more dispersed instrumental distribution compared to both the bootstrap proposal and proposal based on the *extended Kalman filter* (EKF). This broader distribution ensures greater mass allocation to regions with non-negligible target density. Additionally, when the dominant mode switches location, as seen at $t = 1009$,

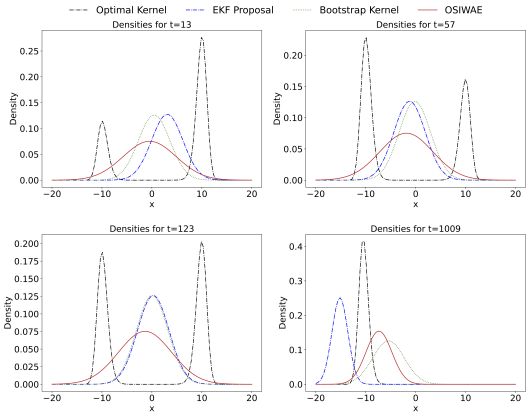

Figure 6: Comparisons between the optimal kernel and the bootstrap, EKF-based, and SMC−OSIWAE-based proposal kernels for the growth model at four distinct timesteps. At $t = 13$, $t = 57$, and $t = 123$ the optimal kernel exhibits two non-negligible modes. At $t = 1009$ there is one unimodal kernel for reference. Despite relying on a univariate proposal, SMC−OSIWAE effectively adjusts the spread of the proposal to better cover the target density across varying regimes compared to the less flexible EKF-based and bootstrap proposals.

SMC−OSIWAE effectively guides the majority of particles towards the new dominant region, further highlighting the advantages of our approach.

6 CONCLUSION

We have introduced OSIWAE, a method for recursively optimising an asymptotic IWAE-type variational objective in SSMs. OSIWAE is equipped with theoretical results describing its objective and its inherent $\mathcal{O}(M^{-1})$ bias with respect to the asymptotic contrast. By using particle methods, we obtain a practically implementable version, SMC−OSIWAE, which can be viewed as an extension of particle-based RML that also allows online training of the particle proposal. Our algorithm also sheds theoretical light on the recently proposed OVSMC, which lacks theoretical underpinnings due to the *ad hoc* truncation of its target gradient. As future research, we intend to provide SMC−OSIWAE with a theoretical analysis akin to that of Tadić and Doucet (2021) for particle-based RML.

Acknowledgements

This work is supported by the Swedish Research Council, grant 2018-05230, and by the Wallenberg AI, Autonomous Systems and Software Program (WASP) *Online learning in dynamical generative models*.

References

Atar, R. and Zeitouni, O. (1997). Exponential stability for nonlinear filtering. *Ann. Inst. H. Poincaré Probab. Statist.*, 33(6):697–725.

Bond-Taylor, S., Leach, A., Long, Y., and Willcocks, C. G. (2021). Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):7327–7347.

Breiman, L. (1960). The strong law of large numbers for a class of Markov chains. *The Annals of Mathematical Statistics*, 31(3):801–803.

Burda, Y., Grosse, R., and Salakhutdinov, R. (2016). Importance weighted autoencoders. In *International Conference on Learning Representations*.

Campbell, A., Shi, Y., Rainforth, T., and Doucet, A. (2021). Online variational filtering and parameter learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 18633–18645. Curran Associates, Inc.

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer, New York.

Chopin, N. and Papaspiliopoulos, O. (2020). *An introduction to sequential Monte Carlo methods*. Springer, New York.

Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A. C., and Bengio, Y. (2015). A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems*, volume 28, pages 2980 – 2988. Curran Associates, Inc.

Cornebise, J., Moulines, E., and Olsson, J. (2008). Adaptive methods for sequential importance sampling with application to state space models. *Stat. Comput.*, 18(4):461–480.

Dau, H.-D. and Chopin, N. (2023). On backward smoothing algorithms. *The Annals of Statistics*, 51(5):2145–2169.

Daudel, K., Benton, J., Shi, Y., and Doucet, A. (2023). Alpha-divergence variational inference meets importance weighted auto-encoders: Methodology and asymptotics. *Journal of Machine Learning Research*, 24(243):1–83.

Del Moral, P., Doucet, A., and Singh, S. S. (2010). A backward interpretation of Feynman-Kac formulae. *ESAIM: Mathematical Modelling and Numerical Analysis*, 44:947–975.

Del Moral, P., Doucet, A., and Singh, S. S. (2015). Uniform stability of a particle approximation of the optimal filter derivative. *SIAM Journal on Control and Optimization*, 53(3):1278–1304.

Dissanayake, G., Newman, P., Clark, S., Durrant-Whyte, H., and Csorba, M. (2001). A solution to the simultaneous localization and map building (SLAM) problem. *IEEE Transactions on Robotics and Automation*, 17(3):229–241.

Douc, R., Garivier, A., Moulines, E., and Olsson, J. (2011). Sequential Monte Carlo smoothing for general state space hidden Markov models. *Ann. Appl. Probab.*, 21(6):1201–2145.

Douc, R. and Matias, C. (2001). Asymptotics of the maximum likelihood estimator for general hidden Markov models. *Bernoulli*, 7(3):381–420.

Doucet, A., De Freitas, N., and Gordon, N., editors (2001). *Sequential Monte Carlo Methods in Practice*. Springer, New York.

Doucet, A., Godsill, S., and Andrieu, C. (2000). On sequential Monte-Carlo sampling methods for Bayesian filtering. *Stat. Comput.*, 10:197–208.

Doucet, A., Moulines, E., and Thin, A. (2023). Differentiable samplers for deep latent variable models. *Philosophical Transactions of the Royal Society A*, 381(2247):20220147.

Finke, A. and Thiery, A. H. (2019). On importance-weighted autoencoders. *arXiv preprint arXiv:1907.10477*.

Gayoso, A., Steier, Z., Lopez, R., Regier, J., Nazor, K. L., Streets, A., and Yosef, N. (2021). Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature methods*, 18(3):272–282.

Gordon, N., Salmond, D., and Smith, A. F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc. F, Radar Signal Process.*, 140:107–113.

Gut, A. (2013). *Probability: A Graduate Course*. Springer Texts in Statistics. Springer New York.

Karimi, B., Miasojedow, B., Moulines, E., and Wai, H.-T. (2019). Non-asymptotic analysis of biased stochastic approximation scheme. In *Conference on Learning Theory*, pages 1944–1974. PMLR.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes. In *International Conference on Learning Representations*.

Kitagawa, G. (1987). Non-Gaussian state space modeling of nonstationary time series. *J. Am. Statist. Assoc.*, 82(400):1023–1063.

Kitagawa, G. and Sato, S. (2001). Monte Carlo smoothing and self-organising state-space model. In *Sequential Monte Carlo methods in practice*, Stat. Eng. Inf. Sci., pages 177–195. Springer, New York.

Le, T. A., Igl, M., Rainforth, T., Jin, T., and Wood, F. (2018). Auto-encoding sequential Monte Carlo. In *International Conference on Learning Representations*.

Le Gland, F. and Mevel, L. (1997). Recursive estimation in HMMs. In *Proc. IEEE Conf. Decis. Control*, pages 3468–3473.

Le Gland, F. and Oudjane, N. (2004). Stability and uniform approximation of nonlinear filters using the Hilbert metric and application to particle filters. *Ann. Appl. Probab.*, 14:144–187.

Maddison, C. J., Lawson, J., Tucker, G., Heess, N., Norouzi, M., Mnih, A., Doucet, A., and Teh, Y. (2017). Filtering variational objectives. *Advances in Neural Information Processing Systems*, 30.

Mastrototaro, A. and Olsson, J. (2024). Online variational sequential Monte Carlo. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 35039–35062. PMLR.

Mastrototaro, A., Olsson, J., and Alenlöv, J. (2024). Fast and numerically stable particle-based online additive smoothing: The adasmooth algorithm. *Journal of the American Statistical Association*, 119(545):356–367.

Meyn, S. P. and Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability*. Cambridge University Press, London.

Naesseth, C., Linderman, S., Ranganath, R., and Blei, D. (2018). Variational sequential Monte Carlo. In *International conference on artificial intelligence and statistics*, pages 968–977. PMLR.

Nowozin, S. (2018). Debiasing evidence approximations: On importance-weighted autoencoders and jackknife variational inference. In *International conference on learning representations*.

Olsson, J., Cappé, O., Douc, R., and Moulines, E. (2008). Sequential Monte Carlo smoothing with application to parameter estimation in non-linear state space models. *Bernoulli*, 14(1):155–179.

Olsson, J. and Westerborn, J. (2017). Efficient particle-based online smoothing in general hidden Markov models: The PaRIS algorithm. *Bernoulli*, 23(3):1951–1996.

Olsson, J. and Westerborn Alenlöv, J. (2020). Particle-based online estimation of tangent filters with application to parameter estimation in nonlinear state-space models. *Annals of the Institute of Statistical Mathematics*, 72:545–576.

Poyiadjis, G., Doucet, A., and Singh, S. S. (2011). Particle approximations of the score and observed information matrix in state space models with application to parameter estimation. *Biometrika*, 98(1):65–80.

Rainforth, T., Kosiorek, A., Le, T. A., Maddison, C., Igl, M., Wood, F., and Teh, Y. W. (2018). Tighter variational bounds are not necessarily better. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4277–4285. PMLR.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Ann. Math. Statist.*, 22:400–407.

Roeder, G., Wu, Y., and Duvenaud, D. K. (2017). Sticking the landing: Simple, lower-variance gradient estimators for variational inference. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Tadić, V. B. and Doucet, A. (2005). Exponential forgetting and geometric ergodicity for optimal filtering in general state-space models. *Stochastic Process. Appl.*, 115(8):1408–1436.

Tadić, V. Z. B. and Doucet, A. (2021). Asymptotic properties of recursive particle maximum likelihood estimation. *IEEE Transactions on Information Theory*, 67(3):1825–1848.

Thrun, S., Burgard, W., and Fox, D. (2005). *Probabilistic Robotics*. MIT Press.

Tucker, G., Lawson, D., Gu, S., and Maddison, C. J. (2018). Doubly reparameterized gradient estimators for monte carlo objectives. In *International Conference on Learning Representations*.

Xu, T., Wang, Y., He, D., Gao, C., Gao, H., Liu, K., and Qin, H. (2022). Multi-sample training for neural image compression. *Advances in Neural Information Processing Systems*, 35:1502–1515.

Zhang, G., Liu, Y., and Jin, X. (2020). A survey of autoencoder-based recommender systems. *Frontiers of Computer Science*, 14:430–450.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Not Applicable]

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Proofs of Section 3.2

In this appendix we present detailed proofs of the results discussed in the main body of the article. We begin with an overview of the structure of the appendix structure to facilitate navigation.

- In Section A.1 we introduce the notation, which includes some measure-theoretic formalism that is not present in the main body of paper.
- In Section A.2 we prove the exponential forgetting of filter and tangent-filter measures (Proposition A.9 and Propostion A.16), as a special case of the results of Tadić and Doucet (2005).
- In Section A.3 we introduce the extended Markov chain, comprising the data generating process, the filter measure associated to the observations and its gradient, and its Markov kernel $\mathbf{T}_\theta$. In addition, for this chain, we establish its ergodicity and a strong law of large numbers for a class of objective functions (Proposition A.22 and Proposition A.23).
- In Section A.4, using the previously established strong law of large number, we define the objective functions $\ell^M(\theta)$ and $\ell(\theta)$ as well as their gradients (Proposition A.27, which proves Proposition 3.1).
- In Section A.5 we study the bias of $\ell^M(\theta)$ and $\nabla\ell^M(\theta)$ with respect to $\ell(\theta)$ and $\nabla\ell(\theta)$ (Corollary A.28 and Corollary A.30, which prove Proposition 3.2 and Theorem 3.3, respectively).
- In Section A.6, we prove some auxiliary lemmas that are used in previous sections.

A.1 Notation

We let $\mathbb{R}_+$ and $\mathbb{R}_+^*$ be the sets of nonnegative and positive real numbers, respectively. For $m \leq n \in \mathbb{N}$, we denote $x_{m:n} := (x_m, x_{m+1}, \ldots, x_{n-1}, x_n)$ or, alternatively, $x^{m:n} := (x^m, x^{m+1}, \ldots, x^{n-1}, x^n)$, depending on the specific case. If $m > n$, then by convention $x_{m:n} = x^{m:n} = \emptyset$, $\prod_{i=m}^n = 1$, and $\sum_{i=m}^n = 0$. For any vector $x_{1:d} \in \mathbb{R}^d$, $d \in \mathbb{N}_{>0}$, we indicate with $\|\cdot\|$ the maximum norm, *i.e.*, $\|x_{1:d}\| = \max_{i\in\{1,\ldots,d\}} |x_i|$. For some general state space $(\mathsf{S}, \mathcal{S})$ we let $\mathsf{F}(\mathcal{S})$ be the set of real Borel-measurable functions on S and $\mathbb{1}_\mathsf{S} \in \mathsf{F}(\mathsf{S})$ be the constant function equal to one on the whole S. We let $\mathsf{M}(\mathcal{S})$ be the set of finite measures on $\mathcal{S}$ and $\mathsf{M}_1(\mathcal{S}) \subset \mathsf{M}(\mathcal{S})$ the set of probability measures. The set of finite signed measures on $\mathcal{S}$ is denoted by $\tilde{\mathsf{M}}(\mathcal{S}) \supset \mathsf{M}(\mathcal{S})$. For $\mu \in \tilde{\mathsf{M}}(\mathcal{S})$, we denote by $|\mu| = \mu^+ + \mu^-$ its total variation, where $\mu^+ \in \mathsf{M}(\mathcal{S})$ and $\mu^- \in \mathsf{M}(\mathcal{S})$ are the positive and negative parts of μ, respectively, *i.e.*, $\mu = \mu^+ - \mu^-$. We let $\|\mu\|_{\mathsf{TV}} = |\sigma|(\mathsf{S})$ be the total variation norm of μ. We denote by $\mathcal{M}_1(\mathcal{S})$, $\mathcal{M}(\mathcal{S})$ and $\tilde{\mathcal{M}}(\mathcal{S})$ the sigma-fields of $\mathsf{M}_1(\mathcal{S})$, $\mathsf{M}(\mathcal{S})$, and $\tilde{\mathsf{M}}(\mathcal{S})$, respectively, induced by the total variation norm. For every integer $p \in \mathbb{N}_{>0}$, we also define the product space

$$(\tilde{\mathsf{M}}^p(\mathcal{X}), \tilde{\mathcal{M}}^{\otimes p}(\mathcal{X})) = (\underbrace{\tilde{\mathsf{M}}(\mathcal{X}) \times \cdots \times \tilde{\mathsf{M}}(\mathcal{X})}_{p \text{ times}}, \underbrace{\tilde{\mathcal{M}}(\mathcal{X}) \otimes \cdots \otimes \tilde{\mathcal{M}}(\mathcal{X})}_{p \text{ times}}).$$

For every $f \in \mathsf{F}(\mathcal{S})$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p) \in \tilde{\mathsf{M}}^p(\mathcal{S})$, we denote

$$\boldsymbol{\mu} f = \int f(s)\, \boldsymbol{\mu}(ds) = (\mu_1 f, \ldots, \mu_p f) \in \mathbb{R}^p,$$

and, by convention, we still denote with $\|\boldsymbol{\mu}\|_{\mathsf{TV}}$ the maximum norm of the vector of total variation norms, *i.e.*,

$$\|\boldsymbol{\mu}\|_{\mathsf{TV}} = \max_{i\in\{1,\ldots,p\}} \|\mu_i\|_{\mathsf{TV}}\,.$$

Given some state spaces $(\mathsf{S}, \mathcal{S})$, $(\mathsf{S}', \mathcal{S}')$, and $(\mathsf{S}'', \mathcal{S}'')$ and two kernels $\mathbf{K}_1 : \mathsf{S} \times \mathcal{S}' \to \mathbb{R}_+$ and $\mathbf{K}_2 : \mathsf{S}' \times \mathcal{S}'' \to \mathbb{R}_+$, we may define new product Markov kernels by, first, the tensor product $\mathbf{K}_1 \otimes \mathbf{K}_2 : \mathsf{S} \times (\mathcal{S}' \otimes \mathcal{S}'') \to \mathbb{R}_+$ given by, for $s \in \mathsf{S}$ and $f \in \mathsf{F}(\mathcal{S}' \otimes \mathcal{S}'')$,

$$(\mathbf{K}_1 \otimes \mathbf{K}_2) f(s) = \iint f(s', s'')\, \mathbf{K}_1(s, ds')\, \mathbf{K}_2(s', ds'')$$

and, second, the standard product $\mathbf{K}_1\mathbf{K}_2 : \mathsf{S} \times \mathcal{S}'' \to \mathbb{R}_+$ given by, for $s \in \mathsf{S}$ and $f \in \mathsf{F}(\mathcal{S}'')$,

$$\mathbf{K}_1\mathbf{K}_2 f(s) = \iint f(s'')\, \mathbf{K}_1(s, ds')\, \mathbf{K}_2(s', ds'').$$

Similarly, for $\mu \in \mathsf{M}(\mathcal{S})$ and $\mathbf{K} : \mathsf{S} \times \mathcal{S}' \to \mathbb{R}_+$, we define $\mu \otimes \mathbf{K} \in \mathsf{M}(\mathcal{S} \otimes \mathcal{S}')$ and $\mu\mathbf{K} \in \mathsf{M}(\mathcal{S}')$ such that for $f_1 \in \mathsf{F}(\mathcal{S} \otimes \mathcal{S}')$ and $f_2 \in \mathsf{F}(\mathcal{S}')$ we have

$$(\mu \otimes \mathbf{K}) f_1 = \iint f_1(s, s') \mu(ds)\, \mathbf{K}(s, ds'),$$
$$\mu\mathbf{K} f_2 = \int f_2(s') \int \mu(ds)\, \mathbf{K}(s, ds').$$

Moreover, for $\mu \in \mathsf{M}(\mathcal{S})$ and $\mu' \in \mathsf{M}(\mathcal{S}')$, we denote by $\mu \otimes \mu' \in \mathsf{M}(\mathcal{S} \otimes \mathcal{S}')$ the standard measure product given by, for $f \in \mathsf{F}(\mathcal{S} \otimes \mathcal{S}')$,

$$(\mu \otimes \mu') f = \iint f(s, s')\, \mu(ds)\, \mu'(ds')$$

and by $\mu^{\otimes k} \in \mathsf{M}(\mathcal{S}^{\otimes k})$, $k \in \mathbb{N}_{>0}$, the generalised product given by, for $f \in \mathsf{F}(\mathcal{S}^{\otimes k})$,

$$\mu^{\otimes k} f = \int \cdots \int f(s_1, \ldots, s_k) \prod_{m=1}^{k} \mu(ds_m).$$

We assume that all random variables are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and consider state and observation spaces $\mathsf{X} \subseteq \mathbb{R}^{d_x}$ and $\mathsf{Y} \subseteq \mathbb{R}^{d_y}$, respectively, where $(d_x, d_y) \in \mathbb{N}^2_{>0}$. The SSM under consideration is a bivariate Markov chain $(X_t, Y_t)_{t \in \mathbb{N}}$ evolving on $(\mathsf{X} \times \mathsf{Y}, \mathcal{X} \otimes \mathcal{Y})$ according to a dynamics governed by a parametric model, with parameter $\theta = (\theta^1, \ldots, \theta^p) \in \Theta \subseteq \mathbb{R}^p$, $p \in \mathbb{N}_{>0}$, where Θ is some parameter space. The Markov transition kernel of the model is

$$\mathbf{S}_\theta : (\mathsf{X} \times \mathsf{Y}) \times (\mathcal{X} \otimes \mathcal{Y}) \ni ((x, y), A) \mapsto \iint \mathbb{1}_A(x', y')\, \mathbf{M}_\theta(x, dx')\, \mathbf{G}_\theta(x', dy'), \tag{8}$$

where we have introduced the Markov kernels

$$\mathbf{M}_\theta : \mathsf{X} \times \mathcal{X} \ni (x, A) \mapsto \int \mathbb{1}_A(x')\, m_\theta(x, x')\, \lambda_{\mathcal{X}}(dx'),$$
$$\mathbf{G}_\theta : \mathsf{X} \times \mathcal{Y} \ni (x, B) \mapsto \int \mathbb{1}_B(y)\, g_\theta(x, y)\, \lambda_{\mathcal{Y}}(dy),$$

with $m_\theta : \mathsf{X} \times \mathsf{X} \to \mathbb{R}_+$ and $g_\theta : \mathsf{X} \times \mathsf{Y} \to \mathbb{R}_+$ being the state and emission transition densities with respect to the reference measures $\lambda_{\mathcal{X}} \in \mathsf{M}(\mathcal{X})$ and $\lambda_{\mathcal{Y}} \in \mathsf{M}(\mathcal{Y})$. Here we have slightly modified the notation of the main paper, by using the short-hand notation $m_\theta(x, x') = m_\theta(x' \mid x)$ and $g_\theta(x, y) = g_\theta(y \mid x)$, and allowing a general reference measure instead of the Lebesgue measure. The chain is initialised according to $\chi \otimes \mathbf{G}_\theta : \mathcal{X} \otimes \mathcal{Y} \ni A \mapsto \int_A \chi(dx)\, \mathbf{G}_\theta(x, dy)$, where χ is some probability measure on $(\mathsf{X}, \mathcal{X})$ having density $m_0(x)$ with respect to $\lambda_{\mathcal{X}}$.

Given a sequence $(y_t)_{t \geq 0}$ of observations, we define, for each $t \in \mathbb{N}$, the filter measure $\phi_t^\theta \in \mathsf{M}_1(\mathcal{X})$ which satisfies, for every $f \in \mathsf{F}(\mathcal{X})$,

$$\phi_t^\theta f := \frac{\int \cdots \int f(x_t) m_0(x_0) g_\theta(x_0, y_0) \prod_{t'=1}^{t} m_\theta(x_{t'-1}, x_{t'}) g_\theta(x_{t'}, y_{t'})\, \lambda_{\mathcal{X}}(dx_0) \cdots \lambda_{\mathcal{X}}(dx_t)}{\int \cdots \int m_0(x_0) g_\theta(x_0, y_0) \prod_{t'=1}^{t} m_\theta(x_{t'-1}, x_{t'}) g_\theta(x_{t'}, y_{t'})\, \lambda_{\mathcal{X}}(dx_0) \cdots \lambda_{\mathcal{X}}(dx_t)}. \tag{9}$$

The corresponding filter derivative, or tangent filter, is given by $\psi_t^\theta f = \nabla_\theta \phi_t^\theta f$.

We let $\mathbf{R}_\theta : \mathsf{X} \times \mathsf{Y} \times \mathcal{X} \to [0, 1]$ be some *proposal kernel*, parameterised by $\theta \in \Theta$ as well and having transition density $r_\theta : \mathsf{X} \times \mathsf{Y} \times \mathsf{X} \to \mathbb{R}_+$ with respect to $\lambda_{\mathcal{X}}$. This proposal is assumed to be such that for every $(x, y, A) \in \mathsf{X} \times \mathsf{Y} \times \mathcal{X}$,

$$\mathbf{R}_\theta((x, y), A) = 0 \Rightarrow \int \mathbb{1}_A(x') g_\theta(x', y)\, \mathbf{M}_\theta(x, dx') = 0.$$

In order to express the `OSIWAE` samples as explicit differentiable functions of θ, the proposal is assumed to be reparameterisable. More precisely, we assume that there exist some state-space $(\mathsf{U}, \mathcal{U})$, an easily samplable probability measure $\nu \in \mathsf{M}_1(\mathcal{U})$, not depending on θ, and a function $h_\theta : \mathsf{X} \times \mathsf{Y} \times \mathsf{U} \to \mathsf{X}$ such that for all $(x, y) \in \mathsf{X} \times \mathsf{Y}$ and $\theta \in \Theta$, it holds that $\int f(h_\theta(x, y, u))\, \nu(du) = \int f(x')\, \mathbf{R}_\theta((x, y), dx')$ for all bounded real-valued measurable functions f on X; in other words, the pushforward distribution $\nu \circ h_\theta^{-1}(x, y, \cdot)$ coincides with $\mathbf{R}_\theta((x, y), \cdot)$.

On the basis of the proposal kernel, we redefine the reparameterised weight function

$$w_\theta(x, y, u) := \frac{m_\theta(x, h_\theta(x, y, u)) g_\theta(h_\theta(x, y, u), y)}{r_\theta(x, y, h_\theta(x, y, u))},$$

for all $(x, y, u) \in \mathsf{X} \times \mathsf{Y} \times \mathsf{U}$ such that $r_\theta(x, y, h_\theta(x, y, u)) > 0$.

A.2 Exponential forgetting of the filter and its derivative

Our analysis relies on the following assumptions.

Assumption A.1. There exists $\epsilon \in (0,1)$ such that for every $\theta \in \Theta$, $(x,x') \in \mathsf{X}^2$, and $y \in \mathsf{Y}$,

$$\epsilon \leq m_\theta(x,x') \leq \epsilon^{-1}, \qquad \epsilon \leq g_\theta(x,y) \leq \epsilon^{-1}.$$

Under Assumption A.1 we define

$$\varrho_\epsilon := (1-\epsilon^4)/(1+\epsilon^4) \in (0,1). \tag{10}$$

Assumption A.2. There exists $\tilde{\kappa}_1 \in [1\infty)$ such that for every $\theta \in \Theta$, $(x,x') \in \mathsf{X}^2$, and $y \in \mathsf{Y}$,

$$\|\nabla_\theta m_\theta(x,x')\| \vee \|\nabla_\theta g_\theta(x,y)\| \leq \tilde{\kappa}_1.$$

The strong mixing Assumptions A.1–A.2 are standard in the literature and point to applications where the state and parameter spaces are compact.

Remark A.3. Note that by Assumption A.1 it follows that the reference measure $\lambda_{\mathcal{X}}$ is a finite measure on $(\mathsf{X},\mathcal{X})$. Thus, without loss of generality we may assume that $\lambda_{\mathcal{X}}$ is a probability measure.

Definition A.4. For $y \in \mathsf{Y}$, let the Markov kernel $\mathbf{L}_\theta\langle y\rangle : \mathsf{X}\times\mathcal{X} \to \mathbb{R}_+$ and the signed kernel $\tilde{\mathbf{L}}_\theta\langle y\rangle : \mathsf{X}\times\mathcal{X} \to \mathbb{R}^p$ be defined by, for $x \in \mathsf{X}$ and $f \in \mathsf{F}(\mathcal{X})$,

$$\mathbf{L}_\theta\langle y\rangle f(x) := \int f(x') g_\theta(x',y) m_\theta(x,x')\, \lambda_{\mathcal{X}}(dx')$$

and

$$\tilde{\mathbf{L}}_\theta\langle y\rangle f(x) := \int f(x') \nabla_\theta \left\{ g_\theta(x',y) m_\theta(x,x') \right\} \lambda_{\mathcal{X}}(dx').$$

Definition A.5. For $\theta \in \Theta$ let the mappings $\Phi_\theta : \mathsf{M}(\mathcal{X}) \times \mathsf{Y} \to \mathsf{M}(\mathcal{X})$ and $\Psi_\theta : \mathsf{M}(\mathcal{X}) \times \tilde{\mathsf{M}}^p(\mathcal{X}) \times \mathsf{Y} \to \tilde{\mathsf{M}}^p(\mathcal{X})$ be given by

$$\Phi_\theta(\mu,y)f = \frac{\mu \mathbf{L}_\theta\langle y\rangle f}{\mu \mathbf{L}_\theta\langle y\rangle \mathbb{1}_{\mathsf{X}}}$$

and

$$\Psi_\theta(\mu,\tilde{\boldsymbol{\mu}},y)f = \frac{\tilde{\boldsymbol{\mu}}\mathbf{L}_\theta\langle y\rangle f - \tilde{\boldsymbol{\mu}}\mathbf{L}_\theta\langle y\rangle \mathbb{1}_{\mathsf{X}} \Phi_\theta(\mu,y)f + \mu\tilde{\mathbf{L}}_\theta\langle y\rangle f - \mu\tilde{\mathbf{L}}_\theta\langle y\rangle \mathbb{1}_{\mathsf{X}} \Phi_\theta(\mu,y)f}{\mu \mathbf{L}_\theta\langle y\rangle \mathbb{1}_{\mathsf{X}}}. \tag{11}$$

Moreover, for $\theta \in \Theta$ and a sequence $(y_t)_{t\in\mathbb{N}_{>0}}$ in Y we may define recursively, for $t \in \mathbb{N}_{>0}$, the composite mappings $\Phi_\theta^t : \mathsf{M}(\mathcal{X}) \times \mathsf{Y}^t \to \mathsf{M}(\mathcal{X})$ and $\Psi_\theta^t : \mathsf{M}(\mathcal{X}) \times \tilde{\mathsf{M}}(\mathcal{X}) \times \mathsf{Y}^t \to \tilde{\mathsf{M}}^p(\mathcal{X})$ by

$$\begin{aligned}\Phi_\theta^t(\mu, y_{1:t}) &:= \Phi_\theta(\Phi_\theta^{t-1}(\mu, y_{1:t-1}), y_t),\\ \Psi_\theta^t(\mu, \tilde{\boldsymbol{\mu}}, y_{1:t}) &:= \Psi_\theta(\Phi_\theta^{t-1}(\mu, y_{1:t-1}), \Psi_\theta^{t-1}(\mu, \tilde{\boldsymbol{\mu}}, y_{1:t-1}), y_t).\end{aligned}$$

By convention we let $\Phi_\theta^0(\mu,y) = \Phi_\theta^0(\mu) = \mu$ and $\Psi_\theta^0(\mu,\tilde{\boldsymbol{\mu}},y) = \Psi_\theta^0(\tilde{\boldsymbol{\mu}}) = \tilde{\boldsymbol{\mu}}$.

The following lemma relates the compositions of Φ_θ and Ψ_θ.

Lemma A.6. *Assume that $\mu^\theta \in \mathsf{M}(\mathcal{X})$ is absolutely continuous with respect to the reference measure $\lambda_{\mathcal{X}}$, with a density being differentiable with respect to the parameter θ. Let $\tilde{\boldsymbol{\mu}}^\theta \in \tilde{\mathsf{M}}^p(\mathcal{X})$ be the tangent-filter measure of μ^θ. Then for every $t \in \mathbb{N}_{>0}$, $y_{1:t} \in \mathsf{Y}^t$, and $f \in \mathsf{F}(\mathcal{X})$,*

$$\nabla_\theta \Phi_\theta^t(\mu^\theta, y_{1:t})f = \Psi_\theta^t(\mu^\theta, \tilde{\boldsymbol{\mu}}^\theta, y_{1:t})f.$$

Proof. We proceed by induction, proving first the claim for in the base case $t = 1$. We immediately see that

$$\begin{aligned}\nabla_\theta \Phi_\theta(\mu^\theta, y_1)f &= \nabla_\theta \frac{\mu^\theta \mathbf{L}_\theta\langle y_1\rangle f}{\mu^\theta \mathbf{L}_\theta\langle y_1\rangle \mathbb{1}_{\mathsf{X}}}\\ &= \frac{\tilde{\boldsymbol{\mu}}^\theta \mathbf{L}_\theta\langle y_1\rangle f - \tilde{\boldsymbol{\mu}}^\theta \mathbf{L}_\theta\langle y_1\rangle \mathbb{1}_{\mathsf{X}} \Phi_\theta(\mu^\theta, y_1)f + \mu^\theta \tilde{\mathbf{L}}_\theta\langle y_1\rangle f - \mu^\theta \tilde{\mathbf{L}}_\theta\langle y_1\rangle \mathbb{1}_{\mathsf{X}} \Phi_\theta(\mu^\theta, y_1)f}{\mu^\theta \mathbf{L}_\theta\langle y_1\rangle \mathbb{1}_{\mathsf{X}}} = \Psi_\theta(\mu^\theta, \tilde{\boldsymbol{\mu}}^\theta, y_1)f.\end{aligned} \tag{12}$$

Now assume that the claim is true for some $t \in \mathbb{N}_{>0}$ and let $\mu_t^\theta = \Phi_\theta^t(\mu^\theta, y_{1:t})$ and $\tilde{\boldsymbol{\mu}}_t^\theta = \Psi_\theta^t(\mu^\theta, \tilde{\boldsymbol{\mu}}^\theta, y_{1:t})$, so that, by the induction hypothesis, $\nabla_\theta \mu_t^\theta f = \tilde{\boldsymbol{\mu}}_t^\theta f$. Thus,

$$\nabla_\theta \Phi_\theta^{t+1}(\mu^\theta, y_{1:t+1}) f = \nabla_\theta \Phi_\theta(\mu_t^\theta, y_{t+1}) f = \Psi_\theta(\mu_t^\theta, \tilde{\boldsymbol{\mu}}_t^\theta, y_{t+1}) f,$$

where we used, first, the recursive definition for Φ_θ^{t+1}, then the induction hypothesis together with (12). The proof is completed by noting that

$$\Psi_\theta(\mu_t^\theta, \tilde{\boldsymbol{\mu}}_t^\theta, y_{t+1}) f = \Psi_\theta(\Phi_\theta^t(\mu^\theta, y_{1:t}), \Psi_\theta^t(\mu^\theta, \tilde{\boldsymbol{\mu}}^\theta, y_{1:t}), y_{t+1}) f = \Psi_\theta^{t+1}(\mu^\theta, \tilde{\boldsymbol{\mu}}^\theta, y_{1:t+1}) f.$$

□

Remark A.7. An immediate consequence of Lemma A.6 is that for $t \in \mathbb{N}$, it holds $\phi_t^\theta = \Phi_\theta(\phi_{t-1}^\theta, y_t) = \Phi_\theta^t(\phi_0^\theta, y_{1:t})$ and $\psi_t^\theta = \Psi_\theta(\phi_{t-1}^\theta, \psi_{t-1}^\theta, y_t) = \Psi_\theta^t(\phi_0^\theta, \psi_0^\theta, y_{1:t})$.

In the following, let d be the Hilbert distance between elements in $\mathsf{M}(\mathcal{X})$, defined as

$$d(\mu, \mu') := \log \frac{\sup_{B \in \mathcal{X} : \mu'(B) > 0} \mu(B)/\mu'(B)}{\inf_{B \in \mathcal{X} : \mu'(B) > 0} \mu(B)/\mu'(B)},$$

where is it assumed there exist $0 < a \leq b$ such that $a\mu(B) \leq \mu'(B) \leq b\mu(B)$ for all $B \in \mathcal{X}$; otherwise $d(\mu, \mu') = \infty$.

Lemma A.8. *Let Assumption A.1 hold. Then for every $(\mu, \mu') \in \mathsf{M}_1(\mathcal{X})^2$, $y \in \mathsf{Y}$, and $\theta \in \Theta$,*

$$\|\mu - \mu'\|_{\mathsf{TV}} \leq \frac{2}{\log 3} d(\mu, \mu'), \tag{13}$$

$$d(\mu \mathbf{L}_\theta \langle y \rangle, \mu' \mathbf{L}_\theta \langle y \rangle) \leq \epsilon^{-4} \|\mu - \mu'\|_{\mathsf{TV}}, \tag{14}$$

$$d(\mu \mathbf{L}_\theta \langle y \rangle, \mu' \mathbf{L}_\theta \langle y \rangle) \leq \frac{1 - \epsilon^4}{1 + \epsilon^4} d(\mu, \mu'). \tag{15}$$

Proof. The proof of (13) can be found in (Atar and Zeitouni, 1997, Lemma 1), while the proofs of (14) and (15) can be found in (Le Gland and Oudjane, 2004, Lemma 3.4 and Proposition 3.9(i)). □

We are now ready to state a first—now classical—result on the exponential forgetting of the filter.

Proposition A.9 (Forgetting of the filter)**.** *Let Assumption A.1 hold. Then there exists $\kappa_1 > 1$, depending on ϵ only, such that for every $t \in \mathbb{N}$, $y_{1:t} \in \mathsf{Y}^t$, $\theta \in \Theta$, and $(\mu, \mu') \in \mathsf{M}_1(\mathcal{X})^2$,*

$$\left\| \Phi_\theta^t(\mu, y_{1:t}) - \Phi_\theta^t(\mu', y_{1:t}) \right\|_{\mathsf{TV}} \leq \kappa_1 \varrho_\epsilon^t \|\mu - \mu'\|_{\mathsf{TV}},$$

where ϱ_ϵ is defined in (10)

Proof. For all $t \in \mathbb{N}$, let $\mu_t := \Phi_\theta^t(\mu, y_{1:t})$ and $\mu_t' := \Phi_\theta^t(\mu', y_{1:t})$. Note that the Hilbert distance is invariant under multiplication by positive scalars, *i.e.*, $d(\mu_{t+1}, \mu_{t+1}') = d(\mu_t \mathbf{L}_\theta \langle y_{t+1} \rangle, \mu_t' \mathbf{L}_\theta \langle y_{t+1} \rangle)$. Hence, Lemma A.8 implies that

$$\begin{aligned}
\|\mu_t - \mu_t'\|_{\mathsf{TV}} &\leq \frac{2}{\log 3} d(\mu_t, \mu_t'), \\
d(\mu_{t+1}, \mu_{t+1}') &\leq \epsilon^{-4} \|\mu_t - \mu_t'\|_{\mathsf{TV}}, \\
d(\mu_{t+1}, \mu_{t+1}') &\leq \frac{1 - \epsilon^4}{1 + \epsilon^4} d(\mu_t, \mu_t'),
\end{aligned}$$

which in turn implies that

$$\|\mu_t - \mu_t'\|_{\mathsf{TV}} \leq \frac{2}{\log 3} \left(\frac{1 - \epsilon^4}{1 + \epsilon^4} \right)^{t-1} d(\mu_1, \mu_1') \leq \frac{2\epsilon^{-4}}{\log 3} \left(\frac{1 - \epsilon^4}{1 + \epsilon^4} \right)^{t-1} \|\mu - \mu'\|_{\mathsf{TV}}.$$

The proof is now concluded recalling the definition of ϱ_ϵ and letting $\kappa_1 := 2\epsilon^{-4} \varrho_\epsilon^{-1} / \log 3$. □

The following lemmas are instrumental for the proof of the forgetting of the tangent filter established in Proposition A.16.

Lemma A.10. *Let Assumptions A.1–A.2 hold. Then for all $\theta \in \Theta$, $y \in \mathsf{Y}$, and all $\tilde{\mu} \in \tilde{\mathsf{M}}(\mathcal{X})$ such that $\|\tilde{\mu}\|_{\mathsf{TV}} < \infty$,*

$$\left\|\tilde{\mu}\tilde{\mathbf{L}}_\theta\langle y\rangle\right\|_{\mathsf{TV}} \leq 2\tilde{\kappa}_1\epsilon^{-1}\left\|\tilde{\mu}\right\|_{\mathsf{TV}}.$$

Proof. Let μ^+ and μ^- be the positive and negative parts of $\tilde{\mu}$, respectively. Then

$$\left\|\tilde{\mu}\tilde{\mathbf{L}}_\theta\langle y\rangle\right\|_{\mathsf{TV}} \leq \left\|\mu^+\tilde{\mathbf{L}}_\theta\langle y\rangle\right\|_{\mathsf{TV}} + \left\|\mu^-\tilde{\mathbf{L}}_\theta\langle y\rangle\right\|_{\mathsf{TV}}.$$

Using Assumptions A.1–A.2, note that $\mu^\pm\tilde{\mathbf{L}}_\theta\langle y\rangle \in \tilde{\mathsf{M}}^p(\mathcal{X})$ and

$$\left\|\mu^\pm\tilde{\mathbf{L}}_\theta\langle y\rangle\right\|_{\mathsf{TV}} \leq \int \mu^\pm(dx)\ \left(m_\theta(x,x')\left\|\nabla_\theta g_\theta(x',y)\right\| + g_\theta(x',y)\left\|\nabla_\theta m_\theta(x,x')\right\|\right)\lambda_{\mathcal{X}}(dx') \leq 2\tilde{\kappa}_1\epsilon^{-1}\left\|\mu^\pm\right\|_{\mathsf{TV}},$$

which implies

$$\left\|\tilde{\mu}\tilde{\mathbf{L}}_\theta\langle y\rangle\right\|_{\mathsf{TV}} \leq 2\tilde{\kappa}_1\epsilon^{-1}(\left\|\mu^+\right\|_{\mathsf{TV}} + \left\|\mu^-\right\|_{\mathsf{TV}}) = 2\tilde{\kappa}_1\epsilon^{-1}\left\|\tilde{\mu}\right\|_{\mathsf{TV}}.$$

□

Definition A.11. For every $\theta \in \Theta$ and $(y_t)_{t\in\mathbb{N}_{>0}} \in \mathsf{Y}$, define recursively

$$\mathbf{L}_\theta^{t+1}\langle y_{1:t+1}\rangle = \mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbf{L}_\theta\langle y_{t+1}\rangle, \quad t \in \mathbb{N},$$

with the convention $\mathbf{L}_\theta^0\langle y_0\rangle f(x) = f(x)$. In addition, for every $\theta \in \Theta$, define the mappings $\tilde{\mathbf{F}}_\theta^t : \mathsf{M}(\mathcal{X}) \times \tilde{\mathsf{M}}^p(\mathcal{X}) \times \mathsf{Y}^t \to \tilde{\mathsf{M}}^p(\mathcal{X})$, $t \in \mathbb{N}$, and $\tilde{\mathbf{H}}_\theta^t : \mathsf{M}(\mathcal{X}) \times \mathsf{Y}^t \to \tilde{\mathsf{M}}^p(\mathcal{X})$, $t \in \mathbb{N}_{>0}$, by $\tilde{\mathbf{F}}_\theta^0(\mu, \tilde{\boldsymbol{\mu}}, y_0) := \tilde{\boldsymbol{\mu}}$ and, for $f \in \mathsf{F}(\mathcal{X})$,

$$\tilde{\mathbf{F}}_\theta^t(\mu, \tilde{\boldsymbol{\mu}}, y_{1:t})f := \frac{(\tilde{\boldsymbol{\mu}}\mathbf{L}_\theta^t\langle y_{1:t}\rangle - \tilde{\boldsymbol{\mu}}\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}\Phi_\theta^t(\mu, y_{1:t}))f}{\mu\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}}$$

$$\tilde{\mathbf{H}}_\theta^t(\mu, y_{1:t})f := \frac{(\Phi_\theta^{t-1}(\mu, y_{1:t-1})\tilde{\mathbf{L}}_\theta\langle y_t\rangle - \Phi_\theta^{t-1}(\mu, y_{1:t-1})\tilde{\mathbf{L}}_\theta\langle y_t\rangle\mathbb{1}_{\mathsf{X}}\Phi_\theta^t(\mu, y_{1:t}))f}{\Phi_\theta^{t-1}(\mu, y_{1:t-1})\mathbf{L}_\theta\langle y_t\rangle\mathbb{1}_{\mathsf{X}}}.$$

Lemma A.12. *For every $\theta \in \Theta$, $(y_t)_{t\in\mathbb{N}_{>0}} \in \mathsf{Y}$, $\mu \in \mathsf{M}_1(\mathcal{X})$, and $\tilde{\boldsymbol{\mu}} \in \tilde{\mathsf{M}}^p(\mathcal{X})$ it holds that*

$$\Psi_\theta^t(\mu, \tilde{\boldsymbol{\mu}}, y_{1:t}) = \tilde{\mathbf{F}}_\theta^t(\mu, \tilde{\boldsymbol{\mu}}, y_{1:t}) + \sum_{s=1}^t \tilde{\mathbf{F}}_\theta^{t-s}(\mu_s, \tilde{\mathbf{H}}_\theta^s(\mu, y_{1:s}), y_{s+1:t}), \quad t \in \mathbb{N}.$$

Proof. The base case $t = 0$ is trivially true. Now we assume that the claim is true for some $t \in \mathbb{N}$ and proceed by induction. Let $\mu_t = \Phi_\theta^t(\mu, y_{1:t}) \in \mathsf{M}_1(\mathcal{X})$ and $\tilde{\boldsymbol{\mu}}_t = \Psi_\theta^t(\mu, \tilde{\boldsymbol{\mu}}, y_{1:t}) \in \tilde{\mathsf{M}}^p(\mathcal{X})$, for $t \geq 0$. We write

$$\begin{aligned}\tilde{\boldsymbol{\mu}}_{t+1} = \Psi_\theta(\mu_t, \tilde{\boldsymbol{\mu}}_t, y_{t+1}) &= \frac{\tilde{\boldsymbol{\mu}}_t\left(\mathbf{L}_\theta\langle y_{t+1}\rangle - \mathbf{L}_\theta\langle y_{t+1}\rangle\mathbb{1}_{\mathsf{X}}\mu_{t+1}\right)}{\mu_t\mathbf{L}_\theta\langle y_{t+1}\rangle\mathbb{1}_{\mathsf{X}}} + \frac{\mu_t\tilde{\mathbf{L}}_\theta\langle y_{t+1}\rangle - \mu_t\tilde{\mathbf{L}}_\theta\langle y_{t+1}\rangle\mathbb{1}_{\mathsf{X}}\mu_{t+1}}{\mu_t\mathbf{L}_\theta\langle y_{t+1}\rangle\mathbb{1}_{\mathsf{X}}} \\ &= \frac{\tilde{\boldsymbol{\mu}}_t\left(\mathbf{L}_\theta\langle y_{t+1}\rangle - \mathbf{L}_\theta\langle y_{t+1}\rangle\mathbb{1}_{\mathsf{X}}\mu_{t+1}\right)}{\mu_t\mathbf{L}_\theta\langle y_{t+1}\rangle\mathbb{1}_{\mathsf{X}}} + \tilde{\mathbf{H}}_\theta^{t+1}(\mu, y_{1:t+1}),\end{aligned}$$

where we used, first, the recursion (11) first and, second, Definition A.11 of $\tilde{\mathbf{H}}_\theta^t$. Now, we apply the induction hypothesis to $\tilde{\boldsymbol{\mu}}_t$, noticing first that for every $\tilde{\boldsymbol{\nu}} \in \tilde{\mathsf{M}}^p(\mathcal{X})$ and $0 \leq s < t$,

$$\begin{aligned}&\frac{\tilde{\mathbf{F}}_\theta^{t-s}(\mu_s, \tilde{\boldsymbol{\nu}}, y_{s+1:t})\left(\mathbf{L}_\theta\langle y_{t+1}\rangle - \mathbf{L}_\theta\langle y_{t+1}\rangle\mathbb{1}_{\mathsf{X}}\mu_{t+1}\right)}{\mu_t\mathbf{L}_\theta\langle y_{t+1}\rangle\mathbb{1}_{\mathsf{X}}} \\ &\quad= \frac{\left(\tilde{\boldsymbol{\nu}}\mathbf{L}_\theta^{t-s}\langle y_{s+1:t}\rangle - \tilde{\boldsymbol{\nu}}\mathbf{L}_\theta^{t-s}\langle y_{s+1:t}\rangle\mathbb{1}_{\mathsf{X}}\Phi_\theta^{t-s}(\mu_s, y_{s+1:t})\right)\left(\mathbf{L}_\theta\langle y_{t+1}\rangle - \mathbf{L}_\theta\langle y_{t+1}\rangle\mathbb{1}_{\mathsf{X}}\mu_{t+1}\right)}{\mu_s\mathbf{L}_\theta^{t-s}\langle y_{s+1:t}\rangle\mathbb{1}_{\mathsf{X}}\mu_t\mathbf{L}_\theta\langle y_{t+1}\rangle\mathbb{1}_{\mathsf{X}}} \\ &\quad= \frac{\tilde{\boldsymbol{\nu}}\mathbf{L}_\theta^{t+1-s}\langle y_{s+1:t+1}\rangle - \tilde{\boldsymbol{\nu}}\mathbf{L}_\theta^{t+1-s}\langle y_{s+1:t+1}\rangle\mathbb{1}_{\mathsf{X}}\Phi_\theta^{t+1-s}(\mu_s, y_{s+1:t+1})}{\mu_s\mathbf{L}_\theta^{t+1-s}\langle y_{s+1:t+1}\rangle\mathbb{1}_{\mathsf{X}}} \\ &\qquad- \frac{\tilde{\boldsymbol{\nu}}\mathbf{L}_\theta^{t-s}\langle y_{s+1:t}\rangle\mathbb{1}_{\mathsf{X}}(\mu_t\mathbf{L}_\theta\langle y_{t+1}\rangle - \mu_t\mathbf{L}_\theta\langle y_{t+1}\rangle\mathbb{1}_{\mathsf{X}}\mu_{t+1})}{\mu_s\mathbf{L}_\theta^{t-s}\langle y_{s+1:t}\rangle\mathbb{1}_{\mathsf{X}}\mu_t\mathbf{L}_\theta\langle y_{t+1}\rangle\mathbb{1}_{\mathsf{X}}}\end{aligned}$$

$$= \tilde{\mathbf{F}}_\theta^{t+1-s}(\mu_s, \tilde{\boldsymbol{\nu}}, y_{s+1:t+1}) - \frac{\tilde{\boldsymbol{\nu}}\mathbf{L}_\theta^{t-s}\langle y_{s+1:t}\rangle \mathbb{1}_{\mathsf{X}}(\mu_{t+1} - \mu_{t+1})}{\mu_s \mathbf{L}_\theta^{t-s}\langle y_{s+1:t}\rangle \mathbb{1}_{\mathsf{X}}} = \tilde{\mathbf{F}}_\theta^{t+1-s}(\mu_s, \tilde{\boldsymbol{\nu}}, y_{s+1:t+1}).$$

Thus, using the previous relation and the induction step for $\tilde{\boldsymbol{\mu}}_t$ we obtain

$$\begin{aligned}
\tilde{\boldsymbol{\mu}}_{t+1} &= \frac{\tilde{\boldsymbol{\mu}}_t \left(\mathbf{L}_\theta\langle y_{t+1}\rangle - \mathbf{L}_\theta\langle y_{t+1}\rangle \mathbb{1}_{\mathsf{X}}\mu_{t+1}\right)}{\mu_t \mathbf{L}_\theta\langle y_{t+1}\rangle \mathbb{1}_{\mathsf{X}}} + \tilde{\mathbf{H}}_\theta^{t+1}(\mu, y_{1:t+1}) \\
&= \tilde{\mathbf{F}}_\theta^{t}(\mu, \tilde{\boldsymbol{\mu}}, y_{1:t}) \frac{\mathbf{L}_\theta\langle y_{t+1}\rangle - \mathbf{L}_\theta\langle y_{t+1}\rangle \mathbb{1}_{\mathsf{X}}\mu_{t+1}}{\mu_t \mathbf{L}_\theta\langle y_{t+1}\rangle \mathbb{1}_{\mathsf{X}}} \\
&\qquad + \sum_{s=1}^{t} \tilde{\mathbf{F}}_\theta^{t-s}(\mu_s, \tilde{\mathbf{H}}_\theta^s(\mu, y_{1:s}), y_{s+1:t}) \frac{\mathbf{L}_\theta\langle y_{t+1}\rangle - \mathbf{L}_\theta\langle y_{t+1}\rangle \mathbb{1}_{\mathsf{X}}\mu_{t+1}}{\mu_t \mathbf{L}_\theta\langle y_{t+1}\rangle \mathbb{1}_{\mathsf{X}}} + \tilde{\mathbf{H}}_\theta^{t+1}(\mu, y_{1:t+1}) \\
&= \tilde{\mathbf{F}}_\theta^{t+1}(\mu, \tilde{\boldsymbol{\mu}}, y_{1:t+1}) + \sum_{s=1}^{t} \tilde{\mathbf{F}}_\theta^{t+1-s}(\mu_s, \tilde{\mathbf{H}}_\theta^s(\mu, y_{1:s}), y_{s+1:t+1}) + \tilde{\mathbf{H}}_\theta^{t+1}(\mu, y_{1:t+1}) \\
&= \tilde{\mathbf{F}}_\theta^{t+1}(\mu, \tilde{\boldsymbol{\mu}}, y_{1:t+1}) + \sum_{s=1}^{t+1} \tilde{\mathbf{F}}_\theta^{t+1-s}(\mu_s, \tilde{\mathbf{H}}_\theta^s(\mu, y_{1:s}), y_{s+1:t+1}),
\end{aligned}$$

which proves the claim. □

Lemma A.13. *Let Assumption A.1 hold. Then for every $t \in \mathbb{N}$, $y_{1:t} \in \mathsf{Y}^t$, $\theta \in \Theta$, $(\mu, \mu') \in \mathsf{M}_1(\mathcal{X})^2$, and $(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\mu}}') \in \tilde{\mathsf{M}}^p(\mathcal{X})^2$,*

$$\begin{aligned}
&\textit{(i)} \quad \left\|\tilde{\mathbf{F}}_\theta^t(\mu, \tilde{\boldsymbol{\mu}}, y_{1:t})\right\|_{\mathsf{TV}} \leq 2\epsilon^{-4}\kappa_1 \varrho_\epsilon^t \left\|\tilde{\boldsymbol{\mu}}\right\|_{\mathsf{TV}}, \\
&\textit{(ii)} \quad \left\|\tilde{\mathbf{F}}_\theta^t(\mu, \tilde{\boldsymbol{\mu}}, y_{1:t}) - \tilde{\mathbf{F}}_\theta^t(\mu, \tilde{\boldsymbol{\mu}}', y_{1:t})\right\|_{\mathsf{TV}} \leq 2\epsilon^{-4}\kappa_1 \varrho_\epsilon^t \left\|\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}}, \\
&\textit{(iii)} \quad \left\|\tilde{\mathbf{F}}_\theta^t(\mu, \tilde{\boldsymbol{\mu}}, y_{1:t}) - \tilde{\mathbf{F}}_\theta^t(\mu', \tilde{\boldsymbol{\mu}}, y_{1:t})\right\|_{\mathsf{TV}} \leq 2\epsilon^{-8}\kappa_1 \varrho_\epsilon^t \left\|\mu - \mu'\right\|_{\mathsf{TV}} \left\|\tilde{\boldsymbol{\mu}}\right\|_{\mathsf{TV}}.
\end{aligned}$$

Proof. We first assume that $p = 1$, this is, $\tilde{\boldsymbol{\mu}} = \tilde{\mu} \in \tilde{\mathsf{M}}(\mathcal{X})$, which implies that $\tilde{\mathbf{F}}_\theta^t(\mu, \tilde{\mu}, y_{1:t}) \in \tilde{\mathsf{M}}(\mathcal{X})$ as well. Let $\mu_t = \Phi_\theta^t(\mu, y_{1:t})$, and denote by μ^+ and μ^- the positive and negative parts of $\tilde{\mu}$, respectively. If $\|\mu^\pm\|_{\mathsf{TV}} > 0$, then we let $\mu_0^\pm = (\mu^\pm \mathbb{1}_{\mathsf{X}})^{-1}\mu^\pm$ and $\mu_t^\pm = \Phi_\theta^t(\mu_0^\pm, y_{1:t})$; otherwise we let $\{\mu_t^\pm\}_{t\geq 0}$ be a sequence of trivial measures. We start with part (i) and write

$$\begin{aligned}
\tilde{\mathbf{F}}_\theta^t(\mu, \tilde{\mu}, y_{1:t}) &= (\mu \mathbf{L}_\theta^t\langle y_{1:t}\rangle \mathbb{1}_{\mathsf{X}})^{-1}(\tilde{\mu}\mathbf{L}_\theta^t\langle y_{1:t}\rangle - \tilde{\mu}\mathbf{L}_\theta^t\langle y_{1:t}\rangle \mathbb{1}_{\mathsf{X}}\mu_t) \\
&= \frac{(\mu^+\mathbb{1}_{\mathsf{X}}\mu_0^+ - \mu^-\mathbb{1}_{\mathsf{X}}\mu_0^-)\mathbf{L}_\theta^t\langle y_{1:t}\rangle - (\mu^+\mathbb{1}_{\mathsf{X}}\mu_0^+ - \mu^-\mathbb{1}_{\mathsf{X}}\mu_0^-)\mathbf{L}_\theta^t\langle y_{1:t}\rangle \mathbb{1}_{\mathsf{X}}\mu_t}{\mu \mathbf{L}_\theta^t\langle y_{1:t}\rangle \mathbb{1}_{\mathsf{X}}} \\
&= \frac{\mu^+\mathbb{1}_{\mathsf{X}}\mu_0^+\mathbf{L}_\theta^t\langle y_{1:t}\rangle - \mu^-\mathbb{1}_{\mathsf{X}}\mu_0^-\mathbf{L}_\theta^t\langle y_{1:t}\rangle - \mu^+\mathbb{1}_{\mathsf{X}}\mu_0^+\mathbf{L}_\theta^t\langle y_{1:t}\rangle \mathbb{1}_{\mathsf{X}}\mu_t + \mu^-\mathbb{1}_{\mathsf{X}}\mu_0^-\mathbf{L}_\theta^t\langle y_{1:t}\rangle \mathbb{1}_{\mathsf{X}}\mu_t}{\mu \mathbf{L}_\theta^t\langle y_{1:t}\rangle \mathbb{1}_{\mathsf{X}}} \\
&= \frac{\mu^+\mathbb{1}_{\mathsf{X}}\mu_0^+\mathbf{L}_\theta^t\langle y_{1:t}\rangle \mathbb{1}_{\mathsf{X}}\mu_t^+ - \mu^-\mathbb{1}_{\mathsf{X}}\mu_0^-\mathbf{L}_\theta^t\langle y_{1:t}\rangle \mathbb{1}_{\mathsf{X}}\mu_t^-}{\mu \mathbf{L}_\theta^t\langle y_{1:t}\rangle \mathbb{1}_{\mathsf{X}}} - \frac{\mu^+\mathbb{1}_{\mathsf{X}}\mu_0^+\mathbf{L}_\theta^t\langle y_{1:t}\rangle \mathbb{1}_{\mathsf{X}}\mu_t - \mu^-\mathbb{1}_{\mathsf{X}}\mu_0^-\mathbf{L}_\theta^t\langle y_{1:t}\rangle \mathbb{1}_{\mathsf{X}}\mu_t}{\mu \mathbf{L}_\theta^t\langle y_{1:t}\rangle \mathbb{1}_{\mathsf{X}}} \\
&= (\mu \mathbf{L}_\theta^t\langle y_{1:t}\rangle \mathbb{1}_{\mathsf{X}})^{-1} \left(\mu^+\mathbb{1}_{\mathsf{X}}\mu_0^+\mathbf{L}_\theta^t\langle y_{1:t}\rangle \mathbb{1}_{\mathsf{X}}(\mu_t^+ - \mu_t) + \mu^-\mathbb{1}_{\mathsf{X}}\mu_0^-\mathbf{L}_\theta^t\langle y_{1:t}\rangle \mathbb{1}_{\mathsf{X}}(\mu_t - \mu_t^-)\right).
\end{aligned}$$

Now, using Assumption A.1 we obtain that

$$\frac{\mu_0^\pm \mathbf{L}_\theta^t\langle y_{1:t}\rangle \mathbb{1}_{\mathsf{X}}}{\mu \mathbf{L}_\theta^t\langle y_{1:t}\rangle \mathbb{1}_{\mathsf{X}}} = \frac{\mu_0^\pm \mathbf{L}_\theta\langle y_1\rangle \mathbf{L}_\theta^{t-1}\langle y_{2:t}\rangle \mathbb{1}_{\mathsf{X}}}{\mu \mathbf{L}_\theta\langle y_1\rangle \mathbf{L}_\theta^{t-1}\langle y_{2:t}\rangle \mathbb{1}_{\mathsf{X}}} \leq \frac{\epsilon^{-2}\lambda_{\mathcal{X}}\mathbf{L}_\theta^{t-1}\langle y_{2:t}\rangle \mathbb{1}_{\mathsf{X}}}{\epsilon^2\lambda_{\mathcal{X}}\mathbf{L}_\theta^{t-1}\langle y_{2:t}\rangle \mathbb{1}_{\mathsf{X}}} = \epsilon^{-4};$$

thus

$$\begin{aligned}
\left\|\tilde{\mathbf{F}}_\theta^t(\mu, \tilde{\mu}, y_{1:t})\right\|_{\mathsf{TV}} &= \left\|\frac{\mu_0^+\mathbf{L}_\theta^t\langle y_{1:t}\rangle \mathbb{1}_{\mathsf{X}}}{\mu \mathbf{L}_\theta^t\langle y_{1:t}\rangle \mathbb{1}_{\mathsf{X}}}\mu^+\mathbb{1}_{\mathsf{X}}(\mu_t^+ - \mu_t) + \frac{\mu_0^-\mathbf{L}_\theta^t\langle y_{1:t}\rangle \mathbb{1}_{\mathsf{X}}}{\mu \mathbf{L}_\theta^t\langle y_{1:t}\rangle \mathbb{1}_{\mathsf{X}}}\mu^-\mathbb{1}_{\mathsf{X}}(\mu_t - \mu_t^-)\right\|_{\mathsf{TV}} \\
&\leq \epsilon^{-4}\left(\mu^+\mathbb{1}_{\mathsf{X}}\left\|\mu_t^+ - \mu_t\right\|_{\mathsf{TV}} + \mu^-\mathbb{1}_{\mathsf{X}}\left\|\mu_t - \mu_t^-\right\|_{\mathsf{TV}}\right).
\end{aligned}$$

Using Proposition A.9, we obtain

$$\left\|\tilde{\mathbf{F}}_\theta^t(\mu,\tilde{\mu},y_{1:t})\right\|_{\mathsf{TV}} \le \epsilon^{-4}\kappa_1\varrho_\epsilon^t\left(\mu^+\mathbb{1}_{\mathsf{X}}\left\|\mu_0^+-\mu_0\right\|_{\mathsf{TV}}+\mu^-\mathbb{1}_{\mathsf{X}}\left\|\mu_0-\mu_0^-\right\|_{\mathsf{TV}}\right)$$
$$\le 2\epsilon^{-4}\kappa_1\varrho_\epsilon^t\left(\mu^+\mathbb{1}_{\mathsf{X}}+\mu^-\mathbb{1}_{\mathsf{X}}\right)=2\epsilon^{-4}\kappa_1\varrho_\epsilon^t\left\|\tilde{\mu}\right\|_{\mathsf{TV}},$$

where we used the fact that the total variation norm of the difference of two probability measures is at most two. In order to generalise to the case $\tilde{\boldsymbol{\mu}}=(\tilde{\mu}_1,\ldots,\tilde{\mu}_p)\in\tilde{\mathsf{M}}^p(\mathcal{X})$, for $p>1$, we simply notice that

$$\left\|\tilde{\mathbf{F}}_\theta^t(\mu,\tilde{\boldsymbol{\mu}},y_{1:t})\right\|_{\mathsf{TV}}=\max_{i\in\{1,\ldots,p\}}\left\|\tilde{\mathbf{F}}_\theta^t(\mu,\tilde{\mu}_i,y_{1:t})\right\|_{\mathsf{TV}}\le 2\epsilon^{-4}\kappa_1\varrho_\epsilon^t\max_{i\in\{1,\ldots,p\}}\left\|\tilde{\mu}_i\right\|_{\mathsf{TV}}=2\epsilon^{-4}\kappa_1\varrho_\epsilon^t\left\|\tilde{\boldsymbol{\mu}}\right\|_{\mathsf{TV}}.$$

Having proved (i), (ii) follows immediately since $\tilde{\mathbf{F}}_\theta^t(\mu,\tilde{\boldsymbol{\mu}},y_{1:t})-\tilde{\mathbf{F}}_\theta^t(\mu,\tilde{\boldsymbol{\mu}}',y_{1:t})=\tilde{\mathbf{F}}_\theta^t(\mu,\tilde{\boldsymbol{\mu}}-\tilde{\boldsymbol{\mu}}',y_{1:t})$. To prove (iii), let $\mu_t=\Phi_\theta^t(\mu,y_{1:t})$ and $\mu_t'=\Phi_\theta^t(\mu',y_{1:t})$ and write

$$\begin{aligned}
&\tilde{\mathbf{F}}_\theta^t(\mu,\tilde{\mu},y_{1:t})-\tilde{\mathbf{F}}_\theta^t(\mu',\tilde{\mu},y_{1:t})\\
&\quad=\frac{\tilde{\mu}\mathbf{L}_\theta^t\langle y_{1:t}\rangle-\tilde{\mu}\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}\mu_t}{\mu\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}}-\frac{\tilde{\mu}\mathbf{L}_\theta^t\langle y_{1:t}\rangle-\tilde{\mu}\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}\mu_t'}{\mu'\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}}\\
&\quad=-\frac{(\mu-\mu')\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}\tilde{\mu}\mathbf{L}_\theta^t\langle y_{1:t}\rangle}{\mu\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}\mu'\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}}-\tilde{\mu}\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}\left(\frac{\mu_t}{\mu\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}}-\frac{\mu_t'}{\mu'\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}}\right)\\
&\quad=-\frac{(\mu-\mu')\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}}{\mu\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}}\frac{\tilde{\mu}\mathbf{L}_\theta^t\langle y_{1:t}\rangle}{\mu'\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}}-\tilde{\mu}\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}\frac{\mu'\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}(\mu_t-\mu_t')-(\mu-\mu')\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}\mu_t'}{\mu\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}\mu'\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}}\\
&\quad=-\frac{(\mu-\mu')\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}}{\mu\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}}\left(\frac{\tilde{\mu}\mathbf{L}_\theta^t\langle y_{1:t}\rangle}{\mu'\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}}-\frac{\tilde{\mu}\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}\mu_t'}{\mu'\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}}\right)-\tilde{\mu}\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}\frac{\mu_t-\mu_t'}{\mu\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}}\\
&\quad=-\frac{(\mu-\mu')\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}}{\mu\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}}\tilde{\mathbf{F}}_\theta^t(\mu',\tilde{\mu},y_{1:t})-\frac{\tilde{\mu}\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}}{\mu\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}}(\mu_t-\mu_t').
\end{aligned}$$

Now, using Assumption A.1 we obtain that

$$\frac{|\tilde{\mu}\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}|}{\mu\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}}=\frac{\left|\tilde{\mu}\mathbf{L}_\theta\langle y_1\rangle\mathbf{L}_\theta^{t-1}\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}\right|}{\mu\mathbf{L}_\theta\langle y_1\rangle\mathbf{L}_\theta^{t-1}\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}}\le\frac{\epsilon^{-2}\left|\tilde{\mu}\mathbb{1}_{\mathsf{X}}\right|\lambda_{\mathcal{X}}\mathbf{L}_\theta^{t-1}\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}}{\epsilon^2\lambda_{\mathcal{X}}\mathbf{L}_\theta^{t-1}\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}}\le\epsilon^{-4}\left\|\tilde{\mu}\right\|_{\mathsf{TV}}$$

and, similarly, $(\mu\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}})^{-1}\left|(\mu-\mu')\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}\right|\le\epsilon^{-4}\left\|\mu-\mu'\right\|_{\mathsf{TV}}$. Therefore,

$$\begin{aligned}
&\left\|\tilde{\mathbf{F}}_\theta^t(\mu,\tilde{\mu},y_{1:t})-\tilde{\mathbf{F}}_\theta^t(\mu',\tilde{\mu},y_{1:t})\right\|_{\mathsf{TV}}\\
&\quad\le\frac{|(\mu-\mu')\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}|}{\mu\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}}\left\|\tilde{\mathbf{F}}_\theta^t(\mu',\tilde{\mu},y_{1:t})\right\|_{\mathsf{TV}}+\frac{|\tilde{\mu}\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}|}{\mu\mathbf{L}_\theta^t\langle y_{1:t}\rangle\mathbb{1}_{\mathsf{X}}}\left\|\mu_t-\mu_t'\right\|_{\mathsf{TV}}\\
&\quad\le\epsilon^{-4}\left(\left\|\mu-\mu'\right\|_{\mathsf{TV}}\left\|\tilde{\mathbf{F}}_\theta^t(\mu',\tilde{\mu},y_{1:t})\right\|_{\mathsf{TV}}+\left\|\tilde{\mu}\right\|_{\mathsf{TV}}\left\|\mu_t-\mu_t'\right\|_{\mathsf{TV}}\right)\\
&\quad\le\epsilon^{-4}\left(\epsilon^{-4}\kappa_1\varrho_\epsilon^t\left\|\tilde{\mu}\right\|_{\mathsf{TV}}\left\|\mu-\mu'\right\|_{\mathsf{TV}}+\kappa_1\varrho_\epsilon^t\left\|\mu-\mu'\right\|_{\mathsf{TV}}\left\|\tilde{\mu}\right\|_{\mathsf{TV}}\right)\\
&\quad\le 2\epsilon^{-8}\kappa_1\varrho_\epsilon^t\left\|\tilde{\mu}\right\|_{\mathsf{TV}}\left\|\mu-\mu'\right\|_{\mathsf{TV}},
\end{aligned}$$

where we in the penultimate inequality used (i) and Proposition A.9. Finally, (iii) is proven letting again $\tilde{\boldsymbol{\mu}}=(\tilde{\mu}_1,\ldots,\tilde{\mu}_p)\in\tilde{\mathsf{M}}^p(\mathcal{X})$, for $p>1$, and writing

$$\begin{aligned}
\left\|\tilde{\mathbf{F}}_\theta^t(\mu,\tilde{\boldsymbol{\mu}},y_{1:t})-\tilde{\mathbf{F}}_\theta^t(\mu',\tilde{\boldsymbol{\mu}},y_{1:t})\right\|_{\mathsf{TV}}&=\max_{i\in\{1,\ldots,p\}}\left\|\tilde{\mathbf{F}}_\theta^t(\mu,\tilde{\mu}_i,y_{1:t})-\tilde{\mathbf{F}}_\theta^t(\mu',\tilde{\mu}_i,y_{1:t})\right\|_{\mathsf{TV}}\\
&\le 2\epsilon^{-8}\kappa_1\varrho_\epsilon^t\left\|\mu-\mu'\right\|_{\mathsf{TV}}\max_{i\in\{1,\ldots,p\}}\left\|\tilde{\mu}_i\right\|_{\mathsf{TV}}\\
&=2\epsilon^{-8}\kappa_1\varrho_\epsilon^t\left\|\mu-\mu'\right\|_{\mathsf{TV}}\left\|\tilde{\boldsymbol{\mu}}\right\|_{\mathsf{TV}}.
\end{aligned}$$

□

Lemma A.14. *Let Assumptions A.1–A.2 hold. Then for every $t\in\mathbb{N}$, $y_{1:t}\in\mathsf{Y}^t$, $\theta\in\Theta$, and $(\mu,\mu')\in\mathsf{M}_1(\mathcal{X})^2$.*

(i) $\left\|\tilde{\mathbf{H}}_\theta^t(\mu,y_{1:t})\right\|_{\mathsf{TV}}\le 4\tilde{\kappa}_1\epsilon^{-3}$,

(ii) $\left\|\tilde{\mathbf{H}}_\theta^t(\mu,y_{1:t})-\tilde{\mathbf{H}}_\theta^t(\mu',y_{1:t})\right\|_{\mathsf{TV}}\le 10\tilde{\kappa}_1\epsilon^{-7}\kappa_1\varrho_\epsilon^{t-1}\left\|\mu-\mu'\right\|_{\mathsf{TV}}$.

Proof. Let $\mu_t = \Phi_\theta^t(\mu, y_{1:t})$ and $\mu'_t = \Phi_\theta^t(\mu', y_{1:t})$. Then, using Definition A.11,

$$\tilde{\mathbf{H}}_\theta^t(\mu, y_{1:t}) = \frac{\mu_{t-1}\tilde{\mathbf{L}}_\theta\langle y_t\rangle - \mu_{t-1}\tilde{\mathbf{L}}_\theta\langle y_t\rangle \mathbb{1}_{\mathsf{X}}\mu_t}{\mu_{t-1}\mathbf{L}_\theta\langle y_t\rangle\mathbb{1}_{\mathsf{X}}},$$

so we may write

$$\left\|\tilde{\mathbf{H}}_\theta^t(\mu, y_{1:t})\right\|_{\mathsf{TV}} \leq \frac{\|\mu_{t-1}\tilde{\mathbf{L}}_\theta\langle y_t\rangle\|_{\mathsf{TV}} + \|\mu_{t-1}\tilde{\mathbf{L}}_\theta\langle y_t\rangle\|_{\mathsf{TV}}\|\mu_t\|_{\mathsf{TV}}}{\mu_{t-1}\mathbf{L}_\theta\langle y_t\rangle\mathbb{1}_{\mathsf{X}}} \leq 2\epsilon^{-2}\left\|\mu_{t-1}\tilde{\mathbf{L}}_\theta\langle y_t\rangle\right\|_{\mathsf{TV}} \leq 4\tilde{\kappa}_1\epsilon^{-3},$$

where we have used Lemma A.10. This establishes (i). In order to prove (ii), let $\mu_t = \Phi_\theta^t(\mu, y_{1:t})$ and $\mu'_t = \Phi_\theta^t(\mu', y_{1:t})$ and write

$$\begin{aligned}
\tilde{\mathbf{H}}_\theta^t(\mu, y_{1:t}) - \tilde{\mathbf{H}}_\theta^t(\mu', y_{1:t}) &= (\mu_{t-1}\tilde{\mathbf{L}}_\theta\langle y_t\rangle - \mu_{t-1}\tilde{\mathbf{L}}_\theta\langle y_t\rangle\mathbb{1}_{\mathsf{X}}\mu_t)\frac{(\mu'_{t-1}-\mu_{t-1})\mathbf{L}_\theta\langle y_t\rangle\mathbb{1}_{\mathsf{X}}}{\mu_{t-1}\mathbf{L}_\theta\langle y_t\rangle\mathbb{1}_{\mathsf{X}}\mu'_{t-1}\mathbf{L}_\theta\langle y_t\rangle\mathbb{1}_{\mathsf{X}}} \\
&\quad + \frac{\mu_{t-1}\tilde{\mathbf{L}}_\theta\langle y_t\rangle - \mu_{t-1}\tilde{\mathbf{L}}_\theta\langle y_t\rangle\mathbb{1}_{\mathsf{X}}\mu_t}{\mu'_{t-1}\mathbf{L}_\theta\langle y_t\rangle\mathbb{1}_{\mathsf{X}}} - \frac{\mu'_{t-1}\tilde{\mathbf{L}}_\theta\langle y_t\rangle - \mu'_{t-1}\tilde{\mathbf{L}}_\theta\langle y_t\rangle\mathbb{1}_{\mathsf{X}}\mu'_t}{\mu'_{t-1}\mathbf{L}_\theta\langle y_t\rangle\mathbb{1}_{\mathsf{X}}} \\
&= (\mu_{t-1}\tilde{\mathbf{L}}_\theta\langle y_t\rangle - \mu_{t-1}\tilde{\mathbf{L}}_\theta\langle y_t\rangle\mathbb{1}_{\mathsf{X}}\mu_t)\frac{(\mu'_{t-1}-\mu_{t-1})\mathbf{L}_\theta\langle y_t\rangle\mathbb{1}_{\mathsf{X}}}{\mu_{t-1}\mathbf{L}_\theta\langle y_t\rangle\mathbb{1}_{\mathsf{X}}\mu'_{t-1}\mathbf{L}_\theta\langle y_t\rangle\mathbb{1}_{\mathsf{X}}} \\
&\quad + \frac{(\mu_{t-1}-\mu'_{t-1})\tilde{\mathbf{L}}_\theta\langle y_t\rangle + (\mu'_{t-1}-\mu_{t-1})\tilde{\mathbf{L}}_\theta\langle y_t\rangle\mathbb{1}_{\mathsf{X}}\mu_t + \mu'_{t-1}\tilde{\mathbf{L}}_\theta\langle y_t\rangle\mathbb{1}_{\mathsf{X}}(\mu'_t-\mu_t)}{\mu'_{t-1}\mathbf{L}_\theta\langle y_t\rangle\mathbb{1}_{\mathsf{X}}}.
\end{aligned}$$

This yields the bound

$$\begin{aligned}
&\left\|\tilde{\mathbf{H}}_\theta^t(\mu, y_{1:t}) - \tilde{\mathbf{H}}_\theta^t(\mu', , y_{1:t})\right\|_{\mathsf{TV}} \\
&\quad\leq \left(\left\|\mu_{t-1}\tilde{\mathbf{L}}_\theta\langle y_t\rangle\right\|_{\mathsf{TV}} + \left\|\mu_{t-1}\tilde{\mathbf{L}}_\theta\langle y_t\rangle\right\|_{\mathsf{TV}}\|\mu_t\|_{\mathsf{TV}}\right)\frac{\left|(\mu'_{t-1}-\mu_{t-1})\mathbf{L}_\theta\langle y_t\rangle\mathbb{1}_{\mathsf{X}}\right|}{\mu_{t-1}\mathbf{L}_\theta\langle y_t\rangle\mathbb{1}_{\mathsf{X}}\mu'_{t-1}\mathbf{L}_\theta\langle y_t\rangle\mathbb{1}_{\mathsf{X}}} \\
&\quad\quad + \frac{\|(\mu_{t-1}-\mu'_{t-1})\tilde{\mathbf{L}}_\theta\langle y_t\rangle\|_{\mathsf{TV}} + \|(\mu'_{t-1}-\mu_{t-1})\tilde{\mathbf{L}}_\theta\langle y_t\rangle\|_{\mathsf{TV}}\|\mu_t\|_{\mathsf{TV}} + \|\mu'_{t-1}\tilde{\mathbf{L}}_\theta\langle y_t\rangle\|_{\mathsf{TV}}\|\mu'_t-\mu_t\|_{\mathsf{TV}}}{\mu'_{t-1}\mathbf{L}_\theta\langle y_t\rangle\mathbb{1}_{\mathsf{X}}} \\
&\quad\leq 4\tilde{\kappa}_1\epsilon^{-1}\frac{\epsilon^{-2}\|\mu'_{t-1}-\mu_{t-1}\|_{\mathsf{TV}}}{\epsilon^4} + \epsilon^{-2}\left(4\tilde{\kappa}_1\epsilon^{-1}\left\|\mu_{t-1}-\mu'_{t-1}\right\|_{\mathsf{TV}} + 2\tilde{\kappa}_1\epsilon^{-1}\|\mu_t-\mu'_t\|_{\mathsf{TV}}\right) \\
&\quad\leq 8\tilde{\kappa}_1\epsilon^{-7}\left\|\mu_{t-1}-\mu'_{t-1}\right\|_{\mathsf{TV}} + 2\tilde{\kappa}_1\epsilon^{-3}\|\mu_t-\mu'_t\|_{\mathsf{TV}} \\
&\quad\leq 8\tilde{\kappa}_1\epsilon^{-7}\kappa_1\varrho_\epsilon^{t-1}\|\mu-\mu'\|_{\mathsf{TV}} + 2\tilde{\kappa}_1\epsilon^{-3}\kappa_1\varrho_\epsilon^t\|\mu-\mu'\|_{\mathsf{TV}} \\
&\quad\leq 10\tilde{\kappa}_1\epsilon^{-7}\kappa_1\varrho_\epsilon^{t-1}\|\mu-\mu'\|_{\mathsf{TV}},
\end{aligned}$$

which finally proves (ii). □

The following lemma, which is a direct consequence of the previous results, will be useful later.

Lemma A.15. *Let Assumptions A.1–A.2 hold. Then for every $\theta \in \Theta$ and $y_0 \in \mathsf{Y}$,*

$$\left\|\psi_0^\theta\right\|_{\mathsf{TV}} \leq 2\epsilon^{-1}\tilde{\kappa}_1.$$

Moreover, there exist constants $c_\Psi > 0$ and $c_\psi > 0$ depending only on ϵ and $\tilde{\kappa}_1$, such that for every $t \in \mathbb{N}$, $y_{1:t} \in \mathsf{Y}^t$, $\theta \in \Theta$, $\mu \in \mathsf{M}_1(\mathcal{X})$, and $\tilde{\boldsymbol{\mu}} \in \tilde{\mathsf{M}}^p(\mathcal{X})$,

$$\left\|\Psi_\theta^t(\mu, \tilde{\boldsymbol{\mu}}, y_{1:t})\right\|_{\mathsf{TV}} \leq c_\Psi(\varrho_\epsilon^t\|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}} + 1).$$

In particular, letting $\psi_t^\theta := \Psi_\theta(\phi_0^\theta, \psi_0^\theta, y_{1:t})$,

$$\left\|\psi_t^\theta\right\|_{\mathsf{TV}} \leq c_\Psi(\varrho_\epsilon^t\left\|\psi_0^\theta\right\|_{\mathsf{TV}} + 1) \leq c_\psi.$$

Proof. Using Lemmas A.12–A.14, we may obtain the bound

$$\left\|\Psi_\theta^t(\mu, \tilde{\boldsymbol{\mu}}, y_{1:t})\right\|_{\mathsf{TV}} \leq \left\|\tilde{\mathbf{F}}_\theta^t(\mu, \tilde{\boldsymbol{\mu}}, y_{1:t})\right\|_{\mathsf{TV}} + \sum_{s=1}^t\left\|\tilde{\mathbf{F}}_\theta^{t-s}(\Phi_\theta(\mu, y_{1:s}), \tilde{\mathbf{H}}_\theta^s(\mu, y_{1:s}), y_{s+1:t})\right\|_{\mathsf{TV}}$$

$$\leq 2\epsilon^{-4}\kappa_1\varrho_\epsilon^t \left\|\tilde{\boldsymbol{\mu}}\right\|_{\mathsf{TV}} + \sum_{s=1}^{t} 2\epsilon^{-4}\kappa_1\varrho_\epsilon^{t-s} \left\|\tilde{\mathbf{H}}_\theta^s(\mu, y_{1:s})\right\|_{\mathsf{TV}}$$
$$\leq 2\epsilon^{-4}\kappa_1\varrho_\epsilon^t \left\|\tilde{\boldsymbol{\mu}}\right\|_{\mathsf{TV}} + 8\epsilon^{-7}\tilde{\kappa}_1\kappa_1 \sum_{s=0}^{t-1} \varrho_\epsilon^s$$
$$\leq 8\epsilon^{-7}\tilde{\kappa}_1\kappa_1(1-\varrho_\epsilon)^{-1}\left(\varrho_\epsilon^t \left\|\tilde{\boldsymbol{\mu}}\right\|_{\mathsf{TV}} + 1\right).$$

Now, for all $\theta \in \Theta$,

$$\left\|\psi_0^\theta\right\|_{\mathsf{TV}} = \int \left\|\nabla_\theta \left(\frac{m_0(x)g_\theta(x,y)}{\int m_0(x')g_\theta(x',y)\,\lambda_{\mathcal{X}}(dx')}\right)\right\| \lambda_{\mathcal{X}}(dx)$$
$$\leq \frac{\int m_0(x)\left\|\nabla_\theta g_\theta(x,y)\right\|\,\lambda_{\mathcal{X}}(dx)}{\int m_0(x')g_\theta(x',y)\,\lambda_{\mathcal{X}}(dx')} + \frac{\int m_0(x)g_\theta(x,y)\,\lambda_{\mathcal{X}}(dx)}{\left(\int m_0(x')g_\theta(x',y)\,\lambda_{\mathcal{X}}(dx')\right)^2}\int m_0(x')\left\|\nabla_\theta g_\theta(x',y)\right\|\,\lambda_{\mathcal{X}}(dx')$$
$$\leq 2\epsilon^{-1}\int m_0(x)\left\|\nabla_\theta g_\theta(x,y)\right\|\,\lambda_{\mathcal{X}}(dx) \leq 2\epsilon^{-1}\tilde{\kappa}_1.$$

Hence,

$$\left\|\psi_t^\theta\right\|_{\mathsf{TV}} = \left\|\Psi_\theta^t(\phi_0^\theta, \psi_0^\theta, y_{1:t})\right\|_{\mathsf{TV}} \leq 8\epsilon^{-7}\tilde{\kappa}_1\kappa_1(1-\varrho_\epsilon)^{-1}(2\epsilon^{-1}\tilde{\kappa}_1+1).$$

The proof is concluded by letting $c_\Psi := 8\epsilon^{-7}\tilde{\kappa}_1\kappa_1(1-\varrho_\epsilon)^{-1}$ and $c_\psi := c_\Psi(2\epsilon^{-1}\tilde{\kappa}_1+1)$. □

Proposition A.16 (exponential forgetting of the tangent filter)**.** *Let Assumptions A.1–A.2 hold. Then there exists $\kappa_2 > 1$, depending only on ϵ and $\tilde{\kappa}_1$, such that for every $t \in \mathbb{N}$, $y_{1:t} \in \mathsf{Y}^t$, $\theta \in \Theta$, $(\mu, \mu') \in \mathsf{M}_1(\mathcal{X})^2$, and $(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\mu}}') \in \tilde{\mathsf{M}}^p(\mathcal{X})^2$,*

$$\left\|\Psi_\theta^t(\mu, \tilde{\boldsymbol{\mu}}, y_{1:t}) - \Psi_\theta^t(\mu', \tilde{\boldsymbol{\mu}}', y_{1:t})\right\|_{\mathsf{TV}} \leq \epsilon^{-4}\kappa_1\varrho_\epsilon^t \left\|\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} + \kappa_2\varrho_\epsilon^t \left\|\mu - \mu'\right\|_{\mathsf{TV}} \left(\left\|\tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} + 1\right)t.$$

Proof. Using Lemma A.12 and Lemma A.13, we may write

$$\begin{aligned}
&\left\|\Psi_\theta^t(\mu, \tilde{\boldsymbol{\mu}}, y_{1:t}) - \Psi_\theta^t(\mu', \tilde{\boldsymbol{\mu}}', y_{1:t})\right\|_{\mathsf{TV}} \\
&\quad\leq \left\|\tilde{\mathbf{F}}_\theta^t(\mu, \tilde{\boldsymbol{\mu}}, y_{1:t}) - \tilde{\mathbf{F}}_\theta^t(\mu', \tilde{\boldsymbol{\mu}}, y_{1:t})\right\|_{\mathsf{TV}} + \left\|\tilde{\mathbf{F}}_\theta^t(\mu', \tilde{\boldsymbol{\mu}}, y_{1:t}) - \tilde{\mathbf{F}}_\theta^t(\mu', \tilde{\boldsymbol{\mu}}', y_{1:t})\right\|_{\mathsf{TV}} \\
&\qquad + \sum_{s=1}^{t} \left\|\tilde{\mathbf{F}}_\theta^{t-s}(\Phi_\theta^s(\mu, y_{1:s}), \tilde{\mathbf{H}}_\theta^s(\mu, y_{1:s}), y_{s:t}) - \tilde{\mathbf{F}}_\theta^{t-s}(\Phi_\theta^s(\mu', y_{1:s}), \tilde{\mathbf{H}}_\theta^s(\mu, y_{1:s}), y_{s:t})\right\|_{\mathsf{TV}} \\
&\qquad + \sum_{s=1}^{t} \left\|\tilde{\mathbf{F}}_\theta^{t-s}(\Phi_\theta^s(\mu', y_{1:s}), \tilde{\mathbf{H}}_\theta^s(\mu, y_{1:s}), y_{s:t}) - \tilde{\mathbf{F}}_\theta^{t-s}(\Phi_\theta^s(\mu', y_{1:s}), \tilde{\mathbf{H}}_\theta^s(\mu', y_{1:s}), y_{s:t})\right\|_{\mathsf{TV}} \\
&\quad\leq 2\epsilon^{-8}\kappa_1\varrho_\epsilon^t \left\|\mu - \mu'\right\|_{\mathsf{TV}} \left\|\tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} + 2\epsilon^{-4}\kappa_1\varrho_\epsilon^t \left\|\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} \\
&\qquad + \sum_{s=1}^{t-1} 2\epsilon^{-8}\kappa_1\varrho_\epsilon^{t-s} \left\|\Phi_\theta^s(\mu, y_{1:s}) - \Phi_\theta^s(\mu', y_{1:s})\right\|_{\mathsf{TV}} \left\|\tilde{\mathbf{H}}_\theta^s(\mu, y_{1:s})\right\|_{\mathsf{TV}} \\
&\qquad + \sum_{s=1}^{t-1} 2\epsilon^{-4}\kappa_1\varrho_\epsilon^{t-s} \left\|\tilde{\mathbf{H}}_\theta^s(\mu, y_{1:s}) - \tilde{\mathbf{H}}_\theta^s(\mu', y_{1:s})\right\|_{\mathsf{TV}} + \left\|\tilde{\mathbf{H}}_\theta^t(\mu, y_{1:t}) - \tilde{\mathbf{H}}_\theta^t(\mu', y_{1:t})\right\|_{\mathsf{TV}}.
\end{aligned}$$

Now, applying Proposition A.9 and Lemma A.14 yields

$$\begin{aligned}
&\left\|\Psi_\theta^t(\mu, \tilde{\boldsymbol{\mu}}, y_{1:t}) - \Psi_\theta^t(\mu', \tilde{\boldsymbol{\mu}}', y_{1:t})\right\|_{\mathsf{TV}} \\
&\leq 2\epsilon^{-8}\kappa_1\varrho_\epsilon^t \left\|\mu - \mu'\right\|_{\mathsf{TV}} \left\|\tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} + \epsilon^{-4}\kappa_1\varrho_\epsilon^t \left\|\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} \\
&\quad + \sum_{s=1}^{t-1} 2\epsilon^{-8}\kappa_1\varrho_\epsilon^{t-s}\kappa_1\varrho_\epsilon^s \left\|\mu - \mu'\right\|_{\mathsf{TV}} 4\tilde{\kappa}_1\epsilon^{-3} \\
&\quad + \sum_{s=1}^{t-1} \epsilon^{-4}\kappa_1\varrho_\epsilon^{t-s} 10\tilde{\kappa}_1\epsilon^{-7}\kappa_1\varrho_\epsilon^{s-1} \left\|\mu - \mu'\right\|_{\mathsf{TV}} + 10\tilde{\kappa}_1\epsilon^{-7}\kappa_1\varrho_\epsilon^{t-1} \left\|\mu - \mu'\right\|_{\mathsf{TV}} \\
&= \epsilon^{-4}\kappa_1\varrho_\epsilon^t \left\|\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} + \left\|\mu - \mu'\right\|_{\mathsf{TV}} \left(2\epsilon^{-8}\kappa_1\varrho_\epsilon^t \left\|\tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} + (t-1)8\epsilon^{-11}\tilde{\kappa}_1\kappa_1\varrho_\epsilon^t\right.
\end{aligned}$$

$$
\begin{aligned}
&\quad +(t-1)10\tilde{\kappa}_1\epsilon^{-11}\kappa_1^2\varrho_\epsilon^{t-1} + 10\tilde{\kappa}_1\epsilon^{-7}\kappa_1\varrho_\epsilon^{t-1}\big) \\
&= \epsilon^{-4}\kappa_1\varrho_\epsilon^t \left\|\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} + 2\epsilon^{-8}\kappa_1\varrho_\epsilon^t \left\|\mu - \mu'\right\|_{\mathsf{TV}} \\
&\qquad \times \left(\left\|\tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} + (t-1)4\epsilon^{-3}\tilde{\kappa}_1 + (t-1)5\tilde{\kappa}_1\epsilon^{-3}\kappa_1\varrho_\epsilon^{-1} + 5\tilde{\kappa}_1\epsilon\varrho_\epsilon^{-1}\right) \\
&\leq \epsilon^{-4}\kappa_1\varrho_\epsilon^t \left\|\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} + 2\epsilon^{-8}\kappa_1\varrho_\epsilon^t \left\|\mu - \mu'\right\|_{\mathsf{TV}} \left(\left\|\tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} + t9\tilde{\kappa}_1\epsilon^{-3}\kappa_1\varrho_\epsilon^{-1}\right) \\
&\leq \epsilon^{-4}\kappa_1\varrho_\epsilon^t \left\|\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} + 18\epsilon^{-11}\kappa_1^2\tilde{\kappa}_1\varrho_\epsilon^{t-1} \left\|\mu - \mu'\right\|_{\mathsf{TV}} \left(\left\|\tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} + 1\right)t.
\end{aligned}
$$

Finally, the proof is completed by defining $\kappa_2 := 18\epsilon^{-11}\kappa_1^2\tilde{\kappa}_1\varrho_\epsilon^{-1}$. □

A.3 Construction and ergodicity of the extended chain

In this section we construct the extended Markov chain, which includes the data-generating SSM, the filter, and the tangent filter, and thus evolves on the product space

$$
(\mathsf{Z}, \mathcal{Z}) := (\mathsf{X} \times \mathsf{Y} \times \mathsf{M}_1(\mathcal{X}) \times \tilde{\mathsf{M}}_0^p(\mathcal{X}), \mathcal{X} \otimes \mathcal{Y} \otimes \mathcal{M}_1(\mathcal{X}) \otimes \tilde{\mathcal{M}}_0^{\otimes p}(\mathcal{X})),
$$

where $(\tilde{\mathsf{M}}_0^p(\mathcal{X}), \tilde{\mathcal{M}}_0^{\otimes p}(\mathcal{X}))$ is the p-fold product of $(\tilde{\mathsf{M}}_0(\mathcal{X}), \tilde{\mathcal{M}}_0(\mathcal{X}))$, the latter being the measurable space of finite signed measures $\tilde{\mu}$ such that $\int \tilde{\mu}(dx) = 0$. In fact, since we will be dealing exclusively with tangent filter measures, this property is always fulfilled. Note that for all $\mu \in \mathsf{M}_1(\mathcal{X})$, $\tilde{\boldsymbol{\mu}} \in \tilde{\mathsf{M}}_0^p(\mathcal{X})$, and $y \in \mathsf{Y}$, it is easily checked that $\Psi_\theta(\mu, \tilde{\boldsymbol{\mu}}, y)\mathbb{1}_{\mathsf{X}} = 0$; thus the tangent filter recursion is zero-mean preserving.

We begin by assuming that the law of the complete data is governed by an unspecified SSM, as stated below.

Assumption A.17. The observed data stream $(Y_t)_{t\in\mathbb{N}}$ is the output of an SSM $(X_t, Y_t)_{t\in\mathbb{N}}$ on $(\mathsf{X} \times \mathsf{Y}, \mathcal{X} \otimes \mathcal{Y})$ with some state and observation transition kernels $\mathbf{M}(x, dx')$ and $\mathbf{G}(x, dy)$, respectively. These kernels have transition densities $m(x, x')$ and $g(x, y)$ with respect to $\lambda_{\mathcal{X}}$ and $\lambda_{\mathcal{Y}}$, respectively. Furthermore, we let $\mathbf{S} : (\mathsf{X} \times \mathsf{Y}) \times (\mathcal{X} \otimes \mathcal{Y}) \to (0, 1)$ be the product kernel $\mathbf{M} \otimes \mathbf{G}$, defined in the same way as its parametric counterpart in (8). The latent transition density satisfies $\epsilon \leq m(x, x') \leq \epsilon^{-1}$ for all $(x, x') \in \mathsf{X}^2$, where ϵ is the same as in Assumption A.1.

Now we introduce the Markov kernel of the extended chain, which is, for every $z = (x_1, y_1, \mu, \tilde{\boldsymbol{\mu}}) \in \mathsf{Z}$ and $f \in \mathsf{F}(\mathcal{Z})$, given by

$$
\mathbf{T}_\theta f(z) = \int f(z')\, \mathbf{T}_\theta(z, dz') = \int f(x_2, y_2, \Phi_\theta(\mu, y_1), \Psi_\theta(\mu, \tilde{\boldsymbol{\mu}}, y_1))\, \mathbf{S}((x_1, y_1), (dx_2, dy_2)).
$$

For $t \in \mathbb{N}_{>0}$, we let $\mathbf{T}_\theta^t$ be the t-skeleton, *i.e.*, $\mathbf{T}_\theta^1 = \mathbf{T}_\theta$ and, recursively, $\mathbf{T}_\theta^{t+1} f(z) = \int \mathbf{T}_\theta^t f(z')\, \mathbf{T}_\theta(z, dz')$ for $z \in \mathsf{Z}$ and $f \in \mathsf{Lip}(\mathcal{Z})$. By convention, we let $\mathbf{T}_\theta^0 f(z) = \delta_z f = f(z)$. It follows that

$$
\mathbf{T}_\theta^t f(z) = \int \cdots \int f(x_{t+1}, y_{t+1}, \Phi_\theta^t(\mu, y_{1:t}), \Psi_\theta^t(\mu, \tilde{\boldsymbol{\mu}}, y_{1:t}))\, \prod_{s=1}^{t} \mathbf{S}((x_s, y_s), (dx_{s+1}, dy_{s+1})).
$$

In the following we will denote by $(Z_t^\theta)_{t\in\mathbb{N}}$ the extended Markov chain governed by the kernel $\mathbf{T}_\theta$, where $Z_t^\theta = (X_{t+1}, Y_{t+1}, \Phi_\theta^t(\phi_0^\theta, Y_{1:t}), \Psi_\theta^t(\phi_0^\theta, \psi_0^\theta, Y_{1:t}))$; note that ϕ_0^θ and ψ_0^θ both depend on the initial observation Y_0. In the following assumption, we formalise how the chain is initialised.

Assumption A.18. The extended chain is initialised by applying a kernel $\boldsymbol{\chi}_\theta : (\mathsf{X} \times \mathsf{Y}) \times \mathcal{Z} \to [0, 1]$ given by, for $(x_0, y_0) \in \mathsf{X} \times \mathsf{Y}$ and $f \in \mathsf{F}(\mathcal{Z})$,

$$
\boldsymbol{\chi}_\theta f(x_0, y_0) := \int f(x_1, y_1, \phi_0^\theta, \psi_0^\theta)\, \mathbf{S}((x_0, y_0), (dx_1, dy_1)),
$$

recalling that ϕ_0^θ and ψ_0^θ are deterministic maps defined in (9). The initial data (X_0, Y_0) is distributed according to $\chi \otimes \mathbf{G}$.

Remark A.19. If we at any point in time $t \in \mathbb{N}_{>0}$ take the conditional expectation of the Markov chain w.r.t. the initial state Z_0^θ, the latter includes information about the first two observed data points $Y_{0:1}$.

Before we can establish the ergodicity of the extended chain, we need the following lemma, which establishes the ergodicity of the data-generating process.

Lemma A.20. *Let Assumption A.17 hold. Then there exists $\sigma \in \mathsf{M}_1(\mathcal{X} \otimes \mathcal{Y})$ such that for all $(x, y) \in \mathsf{X} \times \mathsf{Y}$ and $t \in \mathbb{N}_{>0}$,*

$$
\left\|\mathbf{S}^t((x, y), \cdot) - \sigma\right\|_{\mathsf{TV}} \leq (1 - \epsilon)^t.
$$

Proof. We first note that that the state space $\mathsf{X} \times \mathsf{Y}$ is ν_1-small, where $\nu_1 \in \mathsf{M}(\mathcal{X} \otimes \mathcal{Y})$ is defined as $\nu_1(dx, dy) := \epsilon(\lambda_{\mathcal{X}}(dx) \otimes \mathbf{G}(x, dy))$. In fact, for all $B \in \mathcal{X} \otimes \mathcal{Y}$,

$$\int \mathbb{1}_B(x', y')\, \mathbf{S}((x, y), (dx', dy')) = \int \mathbb{1}_B(x', y')\, \mathbf{M}(x, dx')\, \mathbf{G}(x', dy') \geq \epsilon \int \mathbb{1}_B(x', y')\, \lambda_{\mathcal{X}}(dx')\, \mathbf{G}(x', dy').$$

Then, by Meyn and Tweedie (2009, Theorem 16.2.4(*v*)) it follows that for all $(x, y) \in \mathsf{X} \times \mathsf{Y}$,

$$\left\|\mathbf{S}^t((x, y), \cdot) - \sigma\right\|_{\mathsf{TV}} \leq (1 - \nu_1(\mathsf{X} \times \mathsf{Y}))^t = (1 - \epsilon)^t.$$

□

We now establish a form of uniform geometric ergodicity of $(Z_t^\theta)_{t \in \mathbb{N}}$ for a certain class of measurable functions on Z, which are Lipschitz in the arguments μ and $\tilde{\boldsymbol{\mu}}$.

Definition A.21. Let $\mathsf{Lip}(\mathcal{Z})$ be the set of vector-valued measurable functions on Z, for which there exists a positive constant L_f such that for all $x \in \mathsf{X}$, $y \in \mathsf{Y}$, $(\mu, \mu') \in \mathsf{M}_1(\mathcal{X})^2$, and $(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\mu}}') \in \tilde{\mathsf{M}}_0^p(\mathcal{X})^2$,

(i) $\|f(x, y, \mu, \tilde{\boldsymbol{\mu}})\| \leq L_f(1 + \|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}})$,

(ii) $\left\|f(x, y, \mu, \tilde{\boldsymbol{\mu}}) - f(x, y, \mu', \tilde{\boldsymbol{\mu}}',)\right\| \leq L_f \left(\left\|\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} + (1 + \|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}} + \left\|\tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}}) \|\mu - \mu'\|_{\mathsf{TV}}\right)$,

Proposition A.22 (Uniform ergodicity of the extended Markov chain)**.** *Let Assumptions A.1, A.2, and A.17 hold. Then there exist constants $c > 0$ and $\tilde{c} > 0$, depending on $\epsilon, \tilde{\kappa}_1$ only, such that for all $t \in \mathbb{N}_{>0}$, $\theta \in \Theta$, $(z, z') \in \mathsf{Z}^2$, and $f \in \mathsf{Lip}(\mathcal{Z})$,*

$$\left\|\mathbf{T}_\theta^t f(z) - \mathbf{T}_\theta^t f(z')\right\| \leq c L_f (\|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}} + \left\|\tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} + 1) \varrho_\epsilon^{t/2}$$

and

$$\left\|\mathbf{T}_\theta^{t+1} f(z) - \mathbf{T}_\theta^t f(z)\right\| \leq \tilde{c}(1 - \varrho_\epsilon^{1/2}) L_f (\|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}} + 1) \varrho_\epsilon^{t/2}.$$

Moreover, there exists a kernel $\boldsymbol{\Upsilon}_\theta : \mathsf{Z} \times \mathcal{Z} \to [0, 1]$ such that for every $f \in \mathsf{Lip}(\mathcal{Z})$, $f_\theta := \boldsymbol{\Upsilon}_\theta f(z)$ is a constant and it holds that

$$\left\|\mathbf{T}_\theta^t f(z) - f_\theta\right\| \leq \tilde{c} L_f (\|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}} + 1) \varrho_\epsilon^{t/2}.$$

Proof. Let $z = (x_1, y_1, \mu, \tilde{\boldsymbol{\mu}})$ and $z' = (x_1', y_1', \mu', \tilde{\boldsymbol{\mu}}')$. For every $f \in \mathsf{Lip}(\mathcal{Z})$, we write

$$\begin{aligned}
\mathbf{T}_\theta^t f(z) - \mathbf{T}_\theta^t f(z') = \int \cdots \int \big(& f(x_{t+1}, y_{t+1}, \Phi_\theta^t(\mu, y_{1:t}), \Psi_\theta^t(\mu, \tilde{\boldsymbol{\mu}}, y_{1:t})) \\
& - f(x_{t+1}, y_{t+1}, \Phi_\theta^t(\mu', y_{1:t}), \Psi_\theta^t(\mu', \tilde{\boldsymbol{\mu}}', y_{1:t})) \big) \prod_{s=1}^t \mathbf{S}((x_s, y_s), (dx_{s+1}, dy_{s+1})) \\
& + \int f(x_{t+1}, y_{t+1}, \Phi_\theta^t(\mu', y_{1:t}), \Psi_\theta^t(\mu', \tilde{\boldsymbol{\mu}}', y_{1:t})) \\
& \times (\mathbf{S}((x_1, y_1), (dx_2, dy_2)) - \mathbf{S}((x_1', y_1'), (dx_2, dy_2))) \prod_{s=2}^t \mathbf{S}((x_s, y_s), (dx_{s+1}, dy_{s+1})). \quad (16)
\end{aligned}$$

Let us focus on the first term in (16): by Definition A.21, Lemma A.15, Proposition A.9, and Proposition A.16, it holds that

$$\begin{aligned}
& \left\|f(x_{t+1}, y_{t+1}, \Phi_\theta^t(\mu, y_{1:t}), \Psi_\theta^t(\mu, \tilde{\boldsymbol{\mu}}, y_{1:t})) - f(x_{t+1}, y_{t+1}, \Phi_\theta^t(\mu', y_{1:t}), \Psi_\theta^t(\mu', \tilde{\boldsymbol{\mu}}', y_{1:t}))\right\| \\
& \quad \leq L_f \left\|\Psi_\theta^t(\mu, \tilde{\boldsymbol{\mu}}, y_{1:t}) - \Psi_\theta^t(\mu', \tilde{\boldsymbol{\mu}}', y_{1:t})\right\|_{\mathsf{TV}} \\
& \qquad + L_f (1 + \left\|\Psi_\theta^t(\mu, \tilde{\boldsymbol{\mu}}, y_{1:t})\right\|_{\mathsf{TV}} + \left\|\Psi_\theta^t(\mu', \tilde{\boldsymbol{\mu}}', y_{1:t})\right\|_{\mathsf{TV}}) \left\|\Phi_\theta^t(\mu, y_{1:t}) - \Phi_\theta^t(\mu', y_{1:t})\right\|_{\mathsf{TV}} \\
& \quad \leq L_f (\epsilon^{-4} \kappa_1 \varrho_\epsilon^t \left\|\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} + \kappa_2 \varrho_\epsilon^t \|\mu - \mu'\|_{\mathsf{TV}} (\left\|\tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} + 1) t) \\
& \qquad + L_f (1 + c_\Psi (\varrho_\epsilon^t \|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}} + 1) + c_\Psi (\varrho_\epsilon^t \left\|\tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} + 1)) \kappa_1 \varrho_\epsilon^t \|\mu - \mu'\|_{\mathsf{TV}} \\
& \quad \leq L_f t \varrho_\epsilon^t \left(2\kappa_2 + (1 + 2c_\Psi) 2\kappa_1 + (\epsilon^{-4} \kappa_1 + 2\kappa_1 c_\Psi \varrho_\epsilon^t) \|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}} + (2\kappa_1 c_\Psi \varrho_\epsilon^t + \epsilon^{-4} \kappa_1 + 2\kappa_2) \left\|\tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}}\right) \\
& \quad \leq (2\kappa_2 + 4c_\Psi \kappa_1 + 2\epsilon^{-4} \kappa_1)(1 + \|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}} + \left\|\tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}}) L_f t \varrho_\epsilon^t,
\end{aligned}$$

where we used that $\|\mu - \mu'\|_{\mathsf{TV}} \leq 2$. Now we return to the main expression (16) and write, using Lemma A.20,

$$\begin{aligned}\left\|\mathbf{T}_\theta^t f(z) - \mathbf{T}_\theta^t f(z')\right\| &\leq (2\kappa_2 + 4c_\Psi\kappa_1 + 2\epsilon^{-4}\kappa_1)(1 + \|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}} + \left\|\tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}})L_f t\varrho_\epsilon^t + L_f(1 + \left\|\Psi_\theta^t(\mu', \tilde{\boldsymbol{\mu}}', y_{1:t})\right\|_{\mathsf{TV}}) \\ &\quad \times \left(\int \left|\mathbf{S}^t - \sigma\right|((x_1, dy_1), (x_{t+1}, dy_{t+1})) + \int \left|\mathbf{S}^t - \sigma\right|((x_1', y_1'), (dx_{t+1}, dy_{t+1}))\right) \\ &\leq (2\kappa_2 + 4c_\Psi\kappa_1 + 2\epsilon^{-4}\kappa_1)(1 + \|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}} + \left\|\tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}})L_f t\varrho_\epsilon^t + 2L_f c_\Psi(\varrho_\epsilon^t \left\|\tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} + 1)(1-\epsilon)^t \\ &\leq L_f(1 + \|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}} + \left\|\tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}})\left((2\kappa_2 + 4c_\Psi\kappa_1 + 2\epsilon^{-4}\kappa_1)\sup_{t'\in\mathbb{N}_{>0}} t'\varrho_\epsilon^{t'/2} + 2c_\Psi\right)\varrho_\epsilon^{t/2},\end{aligned}$$

since $\varrho_\epsilon^{1/2} \geq (1-\epsilon)$ and $\sup_{t\in\mathbb{N}_{>0}} t\varrho_\epsilon^{t/2} < \infty$. Letting $c := (2\kappa_2 + 4c_\Psi\kappa_1 + 2\epsilon^{-4}\kappa_1)\sup_{t'\in\mathbb{N}_{>0}} t'\varrho_\epsilon^{t'/2} + 2c_\Psi$ we finally obtain

$$\left\|\mathbf{T}_\theta^t f(z) - \mathbf{T}_\theta^t f(z')\right\| \leq cL_f(\|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}} + \left\|\tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} + 1)\varrho_\epsilon^{t/2}.$$

Now, to prove the second claim, we note that for $t \in \mathbb{N}_{>0}$,

$$\begin{aligned}\left\|\mathbf{T}_\theta^{t+1} f(z) - \mathbf{T}_\theta^t f(z)\right\| &\leq \int \left\|\mathbf{T}_\theta^t f(z') - \mathbf{T}_\theta^t f(z)\right\| \mathbf{T}_\theta(z, dz') \leq cL_f(\|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}} + \int \left\|\tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} \mathbf{T}_\theta(z, dz') + 1)\varrho_\epsilon^{t/2} \\ &= cL_f\left(\|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}} + \|\Psi_\theta(\mu, \tilde{\boldsymbol{\mu}}, y_1)\|_{\mathsf{TV}} + 1\right)\varrho_\epsilon^{t/2}.\end{aligned}$$

By Lemma A.15

$$\|\Psi_\theta(\mu, \tilde{\boldsymbol{\mu}}, y_1)\|_{\mathsf{TV}} \leq c_\Psi\left(\|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}} + 1\right),$$

which implies that

$$\left\|\mathbf{T}_\theta^{t+1} f(z) - \mathbf{T}_\theta^t f(z)\right\| \leq 2cL_f c_\Psi(\|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}} + 1)\varrho_\epsilon^{t/2}.$$

We now define, for $f \in \mathsf{Lip}(\mathcal{Z})$, the kernel

$$\boldsymbol{\Upsilon}_\theta f(z) = \delta_z f + \sum_{t=0}^{\infty}\left(\mathbf{T}_\theta^{t+1} f(z) - \mathbf{T}_\theta^t f(z)\right),$$

for which

$$\begin{aligned}\left\|\mathbf{T}_\theta^t f(z) - \boldsymbol{\Upsilon}_\theta f(z)\right\| = \left\|-\sum_{s=t}^{\infty}\left(\mathbf{T}_\theta^{s+1} f(z) - \mathbf{T}_\theta^s f(z)\right)\right\| &\leq \sum_{s=t}^{\infty}\left\|\mathbf{T}_\theta^{s+1} f(z) - \mathbf{T}_\theta^s f(z)\right\| \\ &\leq 2cL_f c_\Psi(\|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}} + 1)\sum_{s=t}^{\infty}\varrho_\epsilon^{s/2} = \tilde{c}L_f(\|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}} + 1)\varrho_\epsilon^{t/2},\end{aligned}$$

where we have denoted $\tilde{c} := (1 - \varrho_\epsilon^{1/2})^{-1}2c_\Psi c$. Finally, it remains to prove that $\boldsymbol{\Upsilon}_\theta f(z)$ is constant in z. For this purpose, write

$$\begin{aligned}\|\boldsymbol{\Upsilon}_\theta f(z) - \boldsymbol{\Upsilon}_\theta f(z')\| &\leq \inf_{t\in\mathbb{N}_{>0}}\{\left\|\mathbf{T}_\theta^t f(z) - \mathbf{T}_\theta^t f(z')\right\| + \left\|\boldsymbol{\Upsilon}_\theta f(z) - \mathbf{T}_\theta^t f(z)\right\| + \left\|\mathbf{T}_\theta^t f(z') - \boldsymbol{\Upsilon}_\theta f(z')\right\|\} \\ &\leq L_f(c(\|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}} + \left\|\tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} + 1) + \tilde{c}(\|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}} + 1) + \tilde{c}(\left\|\tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} + 1))\inf_{t\in\mathbb{N}_{>0}}\varrho_\epsilon^{t/2} \\ &= 0,\end{aligned}$$

which implies that there exists $f_\theta \in \mathbb{R}^p$ such that $\boldsymbol{\Upsilon}_\theta f(z) = f_\theta$ for all $z \in \mathsf{Z}$. □

Using the previous result, we now establish a strong law of large numbers (LLN) for the given Markov chain.

Proposition A.23. *Let Assumptions A.1, A.2, A.17, and A.18 hold. Then for every $\theta \in \Theta$ and $f \in \mathsf{Lip}(\mathcal{Z})$,*

$$\lim_{t\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[f(Z_t^\theta)\right] = f_\theta$$

and, $\mathbb{P}$-a.s.,

$$\lim_{t\to\infty}\frac{1}{T}\sum_{t=0}^{T-1} f(Z_t^\theta) = f_\theta.$$

Proof. We proceed as in Breiman (1960). First, note that for all $(t, t_0) \in \mathbb{N}^2$, $\mathbb{E}\left[f(Z^\theta_{t_0+t})\middle|Z^\theta_{t_0}\right] = \mathbf{T}^t_\theta f(Z^\theta_{t_0})$. By Lemma A.15,

$$\begin{aligned}\left\|\mathbf{T}^t_\theta f(Z^\theta_{t_0})\right\| &\le \int \|f(z)\|\ \mathbf{T}^t_\theta(Z^\theta_{t_0}, dz) \le L_f \int (1 + \|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}})\, \mathbf{T}^t_\theta(Z^\theta_{t_0}, dz) \\ &= L_f \int \left(1 + \left\|\Psi^{t_0+t}_\theta(\phi^\theta_0, \psi^\theta_0, (Y_{1:t_0+1}, y_{t+2:t_0+t}))\right\|_{\mathsf{TV}}\right)\, \mathbf{S}((X_{t_0+1}, Y_{t_0+1}), (dx_{t_0+2}, dy_{t_0+2})) \\ &\qquad \times \prod_{s=2}^{t} \mathbf{S}((x_{t_0+s}, y_{t_0+s}), (dx_{t_0+s+1}, dy_{t_0+s+1})) \le L_f(1 + c_\psi) < \infty. \quad (17)\end{aligned}$$

Now, by Proposition A.22, for all $\theta \in \Theta$ and $(s_0, t, T) \in \mathbb{N}^3$,

$$\begin{aligned}\left\|\frac{1}{T}\sum_{s=0}^{T-1} \mathbf{T}^{s+s_0}_\theta f(Z^\theta_t) - f_\theta\right\| &\le \frac{1}{T}\sum_{s=0}^{T-1} \left\|\mathbf{T}^{s+s_0}_\theta f(Z^\theta_t) - f_\theta\right\| \\ &\le \frac{1}{T}\tilde{c}L_f\left(\left\|\psi^\theta_t\right\|_{\mathsf{TV}} + 1\right)\varrho_\epsilon^{s_0/2}\sum_{s=0}^{T-1}\varrho_\epsilon^{s/2} \le \frac{1}{T}\tilde{c}L_f(c_\psi + 1)\varrho_\epsilon^{s_0/2}\frac{1 - \varrho_\epsilon^{T/2}}{1 - \varrho_\epsilon^{1/2}}, \quad (18)\end{aligned}$$

where the right-hand-side tends to zero as T tends to infinity. Here we used again Lemma A.15 to bound the total variation of ψ^θ_t. This implies that

$$\lim_{T\to\infty}\frac{1}{T}\sum_{s=0}^{T-1}\mathbb{E}\left[f(Z^\theta_s)\right] = \lim_{T\to\infty}\mathbb{E}\left[\frac{1}{T}\sum_{s=0}^{T-1}\mathbf{T}^s_\theta f(Z^\theta_0)\right] = f_\theta$$

uniformly in θ. Now, define, for $0 \le s < t$,

$$W^{(s)}_t = \mathbf{T}^s_\theta f(Z^\theta_{t-s}) - \mathbf{T}^{s+1}_\theta f(Z^\theta_{t-s-1})$$

and $W^{(s)}_t = 0$ for $s \ge t$; then note that

$$\begin{aligned}\mathbb{E}\left[W^{(s)}_t \mid W^{(s)}_{t-1}, \dots, W^{(s)}_1\right] &= \mathbb{E}\left[\mathbb{E}\left[W^{(s)}_t \mid Z^\theta_{t-s-1}, Z^\theta_{t-s-2}, \dots, Z^\theta_0\right] \mid W^{(s)}_{t-1}, \dots, W^{(s)}_1\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[W^{(s)}_t \mid Z^\theta_{t-s-1}\right] \mid W^{(s)}_{t-1}, \dots, W^{(s)}_1\right] = 0,\end{aligned}$$

where we used, first, the tower property, second, that $(Z^\theta_t)_{t\in\mathbb{N}}$ is a Markov chain, and, third, the fact that

$$\mathbb{E}\left[W^{(s)}_t \mid Z^\theta_{t-s-1}\right] = \mathbf{T}_\theta \mathbf{T}^s_\theta f(Z^\theta_{t-s-1}) - \mathbf{T}^{s+1}_\theta f(Z^\theta_{t-s-1}) = 0.$$

This implies that $(W^{(s)}_t)_{t\in\mathbb{N}}$ is a martingale difference sequence, which, since by (17)

$$\mathbb{E}\left[\|W^{(s)}_t\|^2\right] \le 4L_f^2(1 + c_\psi)^2,$$

satisfies the assumptions of Lemma A.31. Thus, for all $s \in \mathbb{N}$,

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1} W^{(s)}_t = 0, \quad \mathbb{P}\text{-a.s.} \quad (19)$$

Now write

$$f(Z^\theta_t) - \mathbf{T}^{s+1}_\theta f(Z^\theta_{t-s-1}) = \sum_{s'=0}^{s} \mathbf{T}^{s'}_\theta f(Z^\theta_{t-s'}) - \mathbf{T}^{s'+1}_\theta f(Z^\theta_{t-s'-1}) = \sum_{s'=0}^{s} W^{(s')}_t,$$

so that for every $s \in \mathbb{N}$,

$$\left\|\frac{1}{T}\sum_{t=0}^{T-1} f(Z^\theta_t) - \frac{1}{T}\sum_{t=0}^{T-1}\mathbf{T}^{s+1}_\theta f(Z^\theta_t)\right\|$$

$$
\begin{aligned}
&= \left\| \frac{1}{T}\sum_{t=0}^{s} f(Z_t^\theta) + \frac{1}{T}\sum_{t=s+1}^{T-1} f(Z_t^\theta) - \frac{1}{T}\sum_{t=s+1}^{T-1} \mathbf{T}_\theta^{s+1} f(Z_{t-s-1}^\theta) - \frac{1}{T}\sum_{t=T}^{T+s} \mathbf{T}_\theta^{s+1} f(Z_{t-s-1}^\theta) \right\| \\
&\leq \frac{1}{T}\sum_{t=0}^{s} \left\| f(Z_t^\theta) \right\| + \sum_{s'=0}^{s} \left\| \frac{1}{T}\sum_{t=s+1}^{T-1} W_t^{(s')} \right\| + \frac{1}{T}\sum_{t=T}^{T+s} \left\| \mathbf{T}_\theta^{s+1} f(Z_{t-s-1}^\theta) \right\| \\
&\leq \frac{s+1}{T} L_f (1 + c_\psi) + \sum_{s'=0}^{s} \left\| \frac{1}{T}\sum_{t=s+1}^{T-1} W_t^{(s')} \right\| + \frac{s+1}{T} L_f (1 + c_\psi).
\end{aligned}
$$

Letting $T \to \infty$ and using (19), we obtain

$$
\left\| \frac{1}{T}\sum_{t=0}^{T-1} f(Z_t^\theta) - \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{T}_\theta^{s+1} f(Z_t^\theta) \right\| \to 0, \quad \mathbb{P}\text{-a.s.}
$$

Since the previous limit holds for every $s \in \mathbb{N}$, it holds for the average of $S \in \mathbb{N}_{>0}$ elements, *i.e.*,

$$
\lim_{T\to\infty} \left\| \frac{1}{T}\sum_{t=0}^{T-1} f(Z_t^\theta) - \frac{1}{T}\sum_{t=0}^{T-1} \frac{1}{S}\sum_{s=0}^{S-1} \mathbf{T}_\theta^{s+1} f(Z_t^\theta) \right\| \to 0, \quad \mathbb{P}\text{-a.s.}
$$

Finally, for every $S \in \mathbb{N}_{>0}$, we may write

$$
\left\| \frac{1}{T}\sum_{t=0}^{T-1} f(Z_t^\theta) - f_\theta \right\| \leq \left\| \frac{1}{T}\sum_{t=0}^{T-1} f(Z_t^\theta) - \frac{1}{T}\sum_{t=0}^{T-1} \frac{1}{S}\sum_{s=0}^{S-1} \mathbf{T}_\theta^{s+1} f(Z_t^\theta) \right\| + \frac{1}{T}\sum_{t=0}^{T-1} \left\| \frac{1}{S}\sum_{s=0}^{S-1} \mathbf{T}_\theta^{s+1} f(Z_t^\theta) - f_\theta \right\|,
$$

and by (18), we can, for every $\varepsilon > 0$, chose S so large that the right-hand term is smaller than ε. This proves that, $\mathbb{P}$-a.s.,

$$
\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} f(Z_t^\theta) = f_\theta.
$$

□

A.4 Existence of the COLBO

In this section we establish the existence of the COLBO, the contrast function, and their gradients. We begin by letting $z = (x, y, \mu, \tilde{\boldsymbol{\mu}})$ and redefining $V_\theta^M, G_\theta^M, V_\theta$, and G_θ as measurable functions on Z:

$$
\begin{aligned}
V_\theta^M(z) = V_\theta^M(y, \mu) &:= \mathbb{E}_{(\mu\otimes\nu)^{\otimes M}} \left[\log\left(\frac{1}{M}\sum_{i=1}^{M} w_\theta(X^i, y, U^i) \right) \right], \\
G_\theta^M(z) = G_\theta^M(y, \mu, \tilde{\boldsymbol{\mu}}) &:= \mathbb{E}_{(\mu\otimes\nu)^{\otimes M}} \left[\frac{\sum_{i=1}^{M} \nabla_\theta w_\theta(X^i, y, U^i)}{\sum_{i=1}^{M} w_\theta(X^i, y, U^i)} \right] \\
&\quad + M \int \mathbb{E}_{(\mu\otimes\nu)^{\otimes(M-1)}} \left[\log\left(\frac{1}{M} w_\theta(x, y, u) + \frac{1}{M}\sum_{i=1}^{M-1} w_\theta(X^i, y, U^i) \right) \right] \nu(du)\, \tilde{\boldsymbol{\mu}}(dx), \\
V_\theta(z) = V_\theta(y, \mu) &:= \log \int g_\theta(x', y)\, \mathbf{M}_\theta(x, dx')\, \mu(dx), \\
G_\theta(z) = G_\theta(y, \mu, \tilde{\boldsymbol{\mu}}) &:= \frac{\int \nabla_\theta \{ g_\theta(x', y) m_\theta(x, x') \}\, \lambda_{\mathcal{X}}(dx')\, \mu(dx) + \int g_\theta(x', y)\, \mathbf{M}_\theta(x, dx')\, \tilde{\boldsymbol{\mu}}(dx)}{\int g_\theta(x', y)\, \mathbf{M}_\theta(x, dx')\, \mu(dx)}.
\end{aligned}
$$

In the following we establish strong law of large numbers for path averages of $V_\theta^M, G_\theta^M, V_\theta$, and G_θ. This result follows directly from Proposition A.23 if we are able to show that the functions defined above are in $\mathsf{Lip}(\mathcal{Z})$, which requires some additional assumptions listed below.

Assumption A.24. The constants $\epsilon, \tilde{\kappa}_1$ already defined in Assumptions A.1–A.2 satisfy the following additional properties: for all $\theta \in \Theta$, $(x, x') \in \mathsf{X}^2$, $y \in \mathsf{Y}$, and $u \in \mathsf{E}$,

$$\epsilon \leq r_\theta(x, x', y) \leq \epsilon^{-1},$$

$$\max\left\{\left\|\nabla_\theta m_\theta(x, h_\theta(x, y, u))\right\|, \left\|\nabla_\theta g_\theta(h_\theta(x, y, u), y)\right\|, \left\|\nabla_\theta r_\theta(x, h_\theta(x, y, u), y)\right\|\right\} \leq \tilde{\kappa}_1.$$

The following lemma extends these uniform bounds to the weight function.

Lemma A.25. *Let Assumptions A.1–A.2 and A.24 hold. Then $\epsilon^3 \leq w_\theta(x, y, u) \leq \epsilon^{-3}$ and there exists $\tilde{\kappa}_2 \in [1, \infty)$ such that or all $\theta \in \Theta$, $x \in \mathsf{X}$, $y \in \mathsf{Y}$, and $u \in \mathsf{E}$,*

$$\left\|\nabla_\theta w_\theta(x, y, u)\right\| \leq \tilde{\kappa}_2.$$

Proof. The first bound follows immediately from the definition of w_θ and the assumed bounds on m_θ, g_θ, and r_θ. For the latter we write,

$$\begin{aligned}\left\|\nabla_\theta w_\theta(x, y, u)\right\| \leq{}& \frac{g_\theta(h_\theta(x, y, u), y)\left\|\nabla_\theta m_\theta(x, h_\theta(x, y, u))\right\| + m_\theta(x, h_\theta(x, y, u))\left\|\nabla_\theta g_\theta(h_\theta(x, y, u), y)\right\|}{r_\theta(x, h_\theta(x, y, u), y)} \\ &+ \frac{m_\theta(x, h_\theta(x, y, u)) g_\theta(h_\theta(x, y, u), y)}{r_\theta(x, h_\theta(x, y, u), y)^2}\left\|\nabla_\theta r_\theta(x, h_\theta(x, y, u), y)\right\| \leq 2\epsilon^{-2}\tilde{\kappa}_1 + \epsilon^{-4}\tilde{\kappa}_1 =: \tilde{\kappa}_2.\end{aligned}$$

□

We are now ready to prove that $V_\theta^M, G_\theta^M, V_\theta$, and G_θ are all in $\mathsf{Lip}(\mathcal{Z})$.

Lemma A.26. *Let Assumptions A.1, A.2, and A.24 hold. Then for every $\theta \in \Theta$ and $M \geq 2$ it holds that $V_\theta^M, G_\theta^M, V_\theta$, and G_θ are all in $\mathsf{Lip}(\mathcal{Z})$.*

Proof. We begin with V_θ^M, noting that

$$\left|V_\theta^M(z)\right| \leq \mathbb{E}_{(\mu\otimes\nu)^{\otimes M}}\left[\left|\log\left(\frac{1}{M}\sum_{i=1}^M w_\theta(X^i, y, U^i)\right)\right|\right] \leq \log\epsilon^{-3} \leq \log\epsilon^{-3}(1 + \|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}}),$$

since by Lemma A.25, $\epsilon^3 \leq M^{-1}\sum_{i=1}^M w_\theta(X^i, y, U^i) \leq \epsilon^{-3}$. To check condition (ii) of Definition A.21, we write

$$\begin{aligned}\left|V_\theta^M(x, y, \mu, \tilde{\boldsymbol{\mu}}) - V_\theta^M(x, y, \mu', \tilde{\boldsymbol{\mu}}')\right| &\leq \int\left|\log\left(\frac{1}{M}\sum_{i=1}^M w_\theta(x^i, y, u^i)\right)\right|\nu^{\otimes M}(du^{1:M})\left|\mu^{\otimes M} - \mu'^{\otimes M}\right|(dx^{1:M}) \\ &\leq \log\epsilon^{-3}\int\sum_{i=1}^M|\mu - \mu'|(dx^i)\prod_{j=1}^{i-1}\mu(dx^j)\prod_{j=i+1}^M\mu(dx^j) \\ &\leq M\log\epsilon^{-3}\|\mu - \mu'\|_{\mathsf{TV}}.\end{aligned}$$

We continue with G_θ^M, focusing first on the first term of its definition. Using twice Lemma A.25,

$$\left\|\mathbb{E}_{(\mu\otimes\nu)^{\otimes M}}\left[\frac{\sum_{i=1}^M\nabla_\theta w_\theta(X^i, y, U^i)}{\sum_{i=1}^M w_\theta(X^i, y, U^i)}\right]\right\| \leq \mathbb{E}_{(\mu\otimes\nu)^{\otimes M}}\left[\frac{\sum_{i=1}^M\left\|\nabla_\theta w_\theta(X^i, y, U^i)\right\|}{\sum_{i=1}^M w_\theta(X^i, y, U^i)}\right] \leq \epsilon^{-3}\tilde{\kappa}_2$$

and

$$\begin{aligned}&\left\|\mathbb{E}_{(\mu\otimes\nu)^{\otimes M}}\left[\frac{\sum_{i=1}^M\nabla_\theta w_\theta(X^i, y, U^i)}{\sum_{i=1}^M w_\theta(X^i, y, U^i)}\right] - \mathbb{E}_{(\mu'\otimes\nu)^{\otimes M}}\left[\frac{\sum_{i=1}^M\nabla_\theta w_\theta(X^i, y, U^i)}{\sum_{i=1}^M w_\theta(X^i, y, U^i)}\right]\right\| \\ &\leq \int\frac{\sum_{i=1}^M\left\|\nabla_\theta w_\theta(x^i, y, u^i)\right\|}{\sum_{i=1}^M w_\theta(x^i, y, u^i)}\prod_{i=1}^M\nu(du^i)\sum_{j=1}^M|\mu - \mu'|(dx^j)\prod_{i'=1}^{j-1}\mu(dx^{i'})\prod_{i''=j+1}^M\mu'(dx^{i''}) \leq M\epsilon^{-3}\tilde{\kappa}_2\|\mu - \mu'\|_{\mathsf{TV}}.\end{aligned}$$

For the second term of G_θ^M we note that, since $\int\tilde{\boldsymbol{\mu}}(dx) = 0$, for $M \geq 2$,

$$M \int \mathbb{E}_{(\mu\otimes\nu)^{\otimes(M-1)}} \left[\log\left(\frac{1}{M} w_\theta(x,y,u) + \frac{1}{M}\sum_{i=1}^{M-1} w_\theta(X^i,y,U^i)\right)\right] \nu(du)\,\tilde{\boldsymbol{\mu}}(dx)$$
$$= M \int \mathbb{E}_{(\mu\otimes\nu)^{\otimes(M-1)}} \left[\log\left(1 + \frac{w_\theta(x,y,u)}{\sum_{i=1}^{M-1} w_\theta(X^i,y,U^i)}\right)\right] \nu(du)\,\tilde{\boldsymbol{\mu}}(dx)$$
$$+ M\mathbb{E}_{(\mu\otimes\nu)^{\otimes(M-1)}} \left[\log\left(\frac{1}{M}\sum_{i=1}^{M-1} w_\theta(X^i,y,U^i)\right)\right] \int \nu(du) \int \tilde{\boldsymbol{\mu}}(dx),$$

where the second term vanishes. Then taking the norm and using again Lemma A.25, we obtain

$$\left\| M \int \mathbb{E}_{(\mu\otimes\nu)^{\otimes(M-1)}} \left[\log\left(1 + \frac{w_\theta(x,y,u)}{\sum_{i=1}^{M-1} w_\theta(X^i,y,U^i)}\right)\right] \nu(du)\,\tilde{\boldsymbol{\mu}}(dx)\right\|$$
$$\leq M \log\left(\left(1 + \frac{\epsilon^{-6}}{M-1}\right)\right) \|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}} \leq \frac{M\epsilon^{-6}}{M-1} \|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}} \leq 2\epsilon^{-6} \|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}}.$$

It remains to prove that part (ii) of Definition A.21 is satisfied. For this purpose, write

$$\left\| M \int \mathbb{E}_{(\mu\otimes\nu)^{\otimes(M-1)}} \left[\log\left(\frac{1}{M} w_\theta(x,y,u) + \frac{1}{M}\sum_{i=1}^{M-1} w_\theta(X^i,y,U^i)\right)\right] \nu(du)\,\tilde{\boldsymbol{\mu}}(dx) \right.$$
$$\left. - M \int \mathbb{E}_{(\mu'\otimes\nu)^{\otimes(M-1)}} \left[\log\left(\frac{1}{M} w_\theta(x,y,u) + \frac{1}{M}\sum_{i=1}^{M-1} w_\theta(X^i,y,U^i)\right)\right] \nu(du)\,\tilde{\boldsymbol{\mu}}'(dx)\right\|$$
$$\leq \left\| M \int \mathbb{E}_{(\mu\otimes\nu)^{\otimes(M-1)}} \left[\log\left(1 + \frac{w_\theta(x,y,u)}{\sum_{i=1}^{M-1} w_\theta(X^i,y,U^i)}\right)\right] \nu(du)\,(\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}')(dx)\right\|$$
$$+ \int \left| M\mathbb{E}_{(\mu\otimes\nu)^{\otimes(M-1)}} \left[\log\left(1 + \frac{w_\theta(x,y,u)}{\sum_{i=1}^{M-1} w_\theta(X^i,y,U^i)}\right)\right] \right.$$
$$\left. - M\mathbb{E}_{(\mu'\otimes\nu)^{\otimes(M-1)}} \left[\log\left(1 + \frac{w_\theta(x,y,u)}{\sum_{i=1}^{M-1} w_\theta(X^i,y,U^i)}\right)\right]\right| \nu(du)\,\left|\tilde{\boldsymbol{\mu}}'\right|(dx)$$
$$\leq 2\epsilon^{-6} \left\|\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} + 2\epsilon^{-6} M \|\mu - \mu'\|_{\mathsf{TV}} \left\|\tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}}.$$

Finally, we have

$$\left\|G_\theta^M(x,y,\mu,\tilde{\boldsymbol{\mu}})\right\| \leq \epsilon^{-3}\tilde{\kappa}_2 + 2\epsilon^{-6} \|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}} \leq 2\tilde{\kappa}_2\epsilon^{-6}(1 + \|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}})$$

and

$$\left\|G_\theta^M(x,y,\mu,\tilde{\boldsymbol{\mu}}) - G_\theta^M(x,y,\mu',\tilde{\boldsymbol{\mu}}')\right\| \leq M\epsilon^{-3}\tilde{\kappa}_2 \|\mu - \mu'\|_{\mathsf{TV}} + 2\epsilon^{-6} \left\|\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} + 2\epsilon^{-6} M \|\mu - \mu'\|_{\mathsf{TV}} \|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}}$$
$$\leq 2\epsilon^{-6} \left\|\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} + 2M\epsilon^{-6}\tilde{\kappa}_2(1 + \left\|\tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}}) \|\mu - \mu'\|_{\mathsf{TV}}$$
$$\leq 2M\epsilon^{-6}\tilde{\kappa}_2 \left(\left\|\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} + (1 + \|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}} + \left\|\tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}}) \|\mu - \mu'\|_{\mathsf{TV}}\right).$$

For V_θ, we observe that $|V_\theta(x,y,\mu,\tilde{\boldsymbol{\mu}})| \leq \log\epsilon^{-1} \leq \log\epsilon^{-1}(1 + \|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}})$ and

$$\left|V_\theta(x,y,\mu,\tilde{\boldsymbol{\mu}}) - V_\theta(x,y,\mu',\tilde{\boldsymbol{\mu}}')\right| \leq \epsilon^{-1} \int g_\theta(x',y)\,\mathbf{M}_\theta(x,dx')\,\left|\mu - \mu'\right|(dx) \leq \epsilon^{-2} \|\mu - \mu'\|_{\mathsf{TV}}.$$

It remains to show that $G_\theta \in L_{\mathcal{Z}}$. Indeed,

$$\|G_\theta(x,y,\mu,\tilde{\boldsymbol{\mu}})\| \leq \left(\int g_\theta(x',y)\,\mathbf{M}_\theta(x,dx')\,\mu(dx)\right)^{-1} \left(\int g_\theta(x',y)\,\mathbf{M}_\theta(x,dx')\,\left|\tilde{\boldsymbol{\mu}}\right|(dx)\right.$$
$$\left. + \int \left(g_\theta(x',y) \|\nabla_\theta m_\theta(x,x')\| + m_\theta(x,x') \|\nabla_\theta g_\theta(x',y)\|\right) \lambda_{\mathcal{X}}(dx')\,\mu(dx)\right)$$
$$\leq \epsilon^{-1}\left(\epsilon^{-1} \|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}} + \epsilon^{-1}\tilde{\kappa}_1 + \tilde{\kappa}_1\right) \leq 2\epsilon^{-2}\tilde{\kappa}_1(1 + \|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}})$$

and

$$\begin{aligned}\left\|G_\theta(x,y,\mu,\tilde{\boldsymbol{\mu}}) - G_\theta(x,y,\mu',\tilde{\boldsymbol{\mu}}')\right\| &\leq \frac{\int \left\|\nabla_\theta \left(g_\theta(x',y) m_\theta(x,x')\right)\right\| \, \lambda_{\mathcal{X}}(dx') \, |\mu-\mu'|\,(dx)}{\int g_\theta(x',y)\,\mathbf{M}_\theta(x,dx')\,\mu(dx)} \\ &\quad + \frac{\int \left\|\nabla_\theta \left(g_\theta(x',y) m_\theta(x,x')\right)\right\| \, \lambda_{\mathcal{X}}(dx')\,\mu'(dx) \int g_\theta(x',y)\,\mathbf{M}_\theta(x,dx') \, |\mu'-\mu|\,(dx)}{\int g_\theta(x',y)\,\mathbf{M}_\theta(x,dx')\,\mu'(dx) \int g_\theta(x',y)\,\mathbf{M}_\theta(x,dx')\,\mu(dx)} \\ &\quad + \frac{\int g_\theta(x',y)\,\mathbf{M}_\theta(x,dx') \, \left|\tilde{\boldsymbol{\mu}}-\tilde{\boldsymbol{\mu}}'\right|(dx)}{\int g_\theta(x',y)\,\mathbf{M}_\theta(x,dx')\,\mu(dx)} + \frac{\int g_\theta(x',y)\,\mathbf{M}_\theta(x,dx') \, \left|\tilde{\boldsymbol{\mu}}'\right|(dx) \int g_\theta(x',y)\,\mathbf{M}_\theta(x,dx') \, |\mu'-\mu|\,(dx)}{\int g_\theta(x',y)\,\mathbf{M}_\theta(x,dx')\,\mu'(dx) \int g_\theta(x',y)\,\mathbf{M}_\theta(x,dx')\,\mu(dx)} \\ &\leq 2\epsilon^{-2}\tilde{\kappa}_1 \left\|\mu-\mu'\right\|_{\mathsf{TV}} + 2\epsilon^{-4}\tilde{\kappa}_1 \left\|\mu-\mu'\right\|_{\mathsf{TV}} + \epsilon^{-2}\left\|\tilde{\boldsymbol{\mu}}-\tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} + \epsilon^{-4}\left\|\tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} \left\|\mu-\mu'\right\|_{\mathsf{TV}} \\ &\leq 2\epsilon^{-4}\tilde{\kappa}_1 \left(\left\|\tilde{\boldsymbol{\mu}}-\tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}} + (1 + \|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}} + \left\|\tilde{\boldsymbol{\mu}}'\right\|_{\mathsf{TV}}) \left\|\mu-\mu'\right\|_{\mathsf{TV}}\right).\end{aligned}$$

This completes the proof. □

We are finally ready to prove the existence of the contrast function $\ell(\theta)$ and the COLBO $\ell^M(\theta)$ as well as their gradients as $\mathbb{P}$-a.s. limits.

Proposition A.27. *Let Assumptions A.1, A.2, A.17, A.24, and A.18 hold. Then there exist two real-valued differentiable functions ℓ^M and ℓ on Θ such that for every $\theta \in \Theta$ and $M \geq 2$, $\mathbb{P}$-a.s.,*

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} V_\theta^M(Z_t^\theta) = \ell^M(\theta), \qquad \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} G_\theta^M(Z_t^\theta) = \nabla_\theta \ell^M(\theta),$$
$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} V_\theta(Z_t^\theta) = \ell(\theta), \qquad \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} G_\theta(Z_t^\theta) = \nabla_\theta \ell(\theta).$$

Moreover, the same limits hold when the terms of each sum are replaced by their expectations.

Proof. The limits follow from Proposition A.22 and Proposition A.23, respectively, since all functions are in $\mathsf{Lip}(\mathcal{Z})$, as shown in Lemma A.26. It remains to show that the limits for G_θ^M and G_θ are the gradients of $\ell^M(\theta)$ and $\ell(\theta)$, respectively. Indeed, we have

$$\frac{1}{T}\sum_{t=0}^{T-1} \nabla_\theta \mathbb{E}[V_\theta^M(Z_t^\theta)] = \frac{1}{T}\sum_{t=0}^{T-1} \nabla_\theta \mathbb{E}[\mathbf{T}_\theta^t V_\theta^M(Z_0^\theta)] = \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\mathbf{T}_\theta^t G_\theta^M(Z_0^\theta)] = \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[G_\theta^M(Z_t^\theta)], \tag{20}$$

which converges uniformly in θ as $T \to \infty$ by Proposition A.23. The second equality follows from

$$\begin{aligned}\nabla_\theta \mathbb{E}[\mathbf{T}_\theta^t V_\theta^M(Z_0^\theta)] &= \nabla_\theta \int V_\theta^M(x_{t+1}, y_{t+1}, \Phi_\theta^t(\phi_0^\theta, y_{1:t}), \Psi_\theta^t(\phi_0^\theta, \psi_0^\theta, y_{1:t})) \\ &\qquad \times \chi(dx_0)\,\mathbf{G}(x_0, dy_0) \prod_{s=0}^{t} \mathbf{S}((x_s,y_s),(dx_{s+1},dy_{s+1})) \\ &= \int G_\theta^M(x_{t+1}, y_{t+1}, \Phi_\theta^t(\phi_0^\theta, y_{1:t}), \Psi_\theta^t(\phi_0^\theta, \psi_0^\theta, y_{1:t}))\,\chi(dx_0)\,\mathbf{G}(x_0, dy_0) \prod_{s=0}^{t} \mathbf{S}((x_s,y_s),(dx_{s+1},dy_{s+1})) \\ &= \mathbb{E}[\mathbf{T}_\theta^t G_\theta^M(Z_0^\theta)].\end{aligned}$$

Then, since

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[V_\theta^M(Z_t^\theta)] = \ell^M(\theta)$$

uniformly in θ by Proposition A.23, the uniform convergence theorem states that the limit of (20) is $\nabla_\theta \ell^M(\theta)$. The same argument can be applied to G_θ, which concludes the proof. □

A.5 Bias study

In this section we study the bias of the COLBO w.r.t. the contrast function and between the gradient functions G_θ^M and G_θ.

Proposition A.28. *Let Assumptions A.1, A.2, A.17, A.24, and A.18 hold. Then for all $\theta \in \Theta$ and $M \in \mathbb{N}_{>0}$, $\ell(\theta) \geq \ell^{M+1}(\theta) \geq \ell^M(\theta)$. Moreover,*

$$\ell(\theta) - \ell^M(\theta) = \frac{\epsilon^{-8}}{2M} + \mathcal{O}\left(\frac{1}{M^2}\right).$$

Proof. To prove this we use the results of Nowozin (2018). First, we rewrite V_θ as

$$\begin{aligned} V_\theta(z) = V_\theta(x, y, \mu, \tilde{\boldsymbol{\mu}}) &= \log \int g_\theta(x', y) m_\theta(x, x')\, \lambda_{\mathcal{X}}(dx')\, \mu(dx) \\ &= \log \int w_\theta(x, y, u)\, \nu(du)\, \mu(dx) = \log \mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X, y, U)\right], \end{aligned}$$

so that we may express

$$V_\theta(z) - V_\theta^M(z) = \log \mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X, y, U)\right] - \mathbb{E}_{(\mu\otimes\nu)^{\otimes M}}\left[\log\left(\frac{1}{M}\sum_{i=1}^M w_\theta(X^i, y, U^i)\right)\right].$$

Now, since all the moments of $w_\theta(X, y, U)$ are finite, being w_θ bounded, then we may apply Nowozin (2018, Proposition 1). Thus, for all $\theta \in \Theta$ and $z \in \mathsf{Z}$,

$$V_\theta(z) - V_\theta^M(z) = \frac{\mathbb{E}_{\mu\otimes\nu}[(w_\theta(X, y, U))^2]}{2M\left(\mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X, y, U)\right]\right)^2} + \mathcal{O}\left(\frac{1}{M^2}\right) \leq \frac{\epsilon^{-8}}{2M} + \mathcal{O}\left(\frac{1}{M^2}\right).$$

Then, by Proposition A.27 we have

$$\ell(\theta) - \ell^M(\theta) = \lim_{t\to\infty} \frac{1}{t}\sum_{s=0}^{t-1} \mathbb{E}\left[V_\theta(Z_s^\theta) - V_\theta^M(Z_s^\theta)\right] \leq \frac{\epsilon^{-8}}{2M} + \mathcal{O}\left(\frac{1}{M^2}\right). \tag{21}$$

The monotonicity in M follows from (Burda et al., 2016, Theorem 1), which states that

$$V_\theta(z) \geq V_\theta^{M+1}(z) \geq V_\theta^M(z)$$

for all $z \in \mathsf{Z}$ and $M \in \mathbb{N}_{>0}$; then the claim is proven by expressing $\ell(\theta) - \ell^M(\theta)$ and $\ell^{M+1}(\theta) - \ell^M(\theta)$ in the same way as in (21) and using that nonnegative sequences have nonnegative limits. This concludes the proof. □

Theorem A.29. *Let Assumptions A.1, A.2, and A.24 hold. Then, for all $\theta \in \Theta$, $z \in \mathsf{Z}$, and $M > \epsilon^{-6} + 1$,*

$$\left\|G_\theta^M(z) - G_\theta(z)\right\| \leq \frac{\tilde{b}_1}{M} + \frac{(\tilde{b}_2\epsilon^{-3} + 2\epsilon^{-6})\left\|\tilde{\boldsymbol{\mu}}\right\|_{\mathsf{TV}}}{M-1} + 2\epsilon^{-6}\left\|\tilde{\boldsymbol{\mu}}\right\|_{\mathsf{TV}} \sum_{j=2}^\infty \frac{1}{j+1}\left(\frac{\epsilon^{-6}}{M-1}\right)^j,$$

where $\tilde{b}_1$ and $\tilde{b}_2$ are the constants provided in Lemma A.33, depending only on ϵ and $\tilde{\kappa}_2$.

Proof. By Lemma A.32 and Fubini's Theorem we may write

$$\begin{aligned} G_\theta^M(z) &= G_\theta^M(y, \mu, \tilde{\boldsymbol{\mu}}) \\ &= \mathbb{E}_{(\mu\otimes\nu)^{\otimes M}}\left[\frac{\sum_{i=1}^M \nabla_\theta w_\theta(X^i, y, U^i)/M}{\sum_{i=1}^M w_\theta(X^i, y, U^i)/M}\right] + \int \mathbb{E}_{(\mu\otimes\nu)^{\otimes(M-1)}}\left[\frac{M w_\theta(x, y, u)}{\sum_{i=1}^{M-1} w_\theta(X^i, y, U^i)}\right]\nu(du)\,\tilde{\boldsymbol{\mu}}(dx) \\ &\quad + \int M\sum_{j=2}^\infty \frac{(-1)^{j+1}}{j}\mathbb{E}_{(\mu\otimes\nu)^{\otimes(M-1)}}\left[\left(\frac{w_\theta(x, y, u)}{\sum_{i=1}^{M-1} w_\theta(X^i, y, U^i)}\right)^j\right]\nu(du)\,\tilde{\boldsymbol{\mu}}(dx). \end{aligned} \tag{22}$$

In order to be able to compare G_θ to G_θ^M, we express the former as

$$G_\theta(z) = G_\theta(x,y,\mu,\tilde{\boldsymbol{\mu}}) = \frac{\nabla_\theta \int g_\theta(x',y)\,\mathbf{M}_\theta(x,dx')\,\mu(dx) + \int g_\theta(x',y)\,\mathbf{M}_\theta(x,dx')\,\tilde{\boldsymbol{\mu}}(dx)}{\int g_\theta(x',y)\,\mathbf{M}_\theta(x,dx')\,\mu(dx)}$$
$$= \frac{\int \nabla_\theta w_\theta(x,y,u)\,\nu(du)\,\mu(dx) + \int w_\theta(x,y,u)\,\nu(du)\,\tilde{\boldsymbol{\mu}}(dx)}{\int w_\theta(x,y,u)\,\nu(du)\,\mu(dx)}$$
$$= \frac{\mathbb{E}_{\mu\otimes\nu}\left[\nabla_\theta w_\theta(X,y,U)\right]}{\mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right]} + \frac{\int w_\theta(x,y,u)\nu(du)\,\tilde{\boldsymbol{\mu}}(dx)}{\mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right]}. \quad (23)$$

To bound the bias, we begin with the first terms of (22) and (23), and by Lemma A.33,

$$\left\| \frac{\mathbb{E}_{\mu\otimes\nu}\left[\nabla_\theta w_\theta(X,y,U)\right]}{\mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right]} - \mathbb{E}_{(\mu\otimes\nu)^{\otimes M}}\left[\frac{\sum_{i=1}^M \nabla_\theta w_\theta(X^i,y,U^i)/M}{\sum_{i=1}^M w_\theta(X^i,y,U^i)/M}\right] \right\| \le \frac{\tilde{b}_1}{M}. \quad (24)$$

The second term of (22) is compared to the second term of (23), and again by Lemma A.33,

$$\left\| \int \left(\mathbb{E}_{(\mu\otimes\nu)^{\otimes(M-1)}}\left[\frac{w_\theta(x,y,u)M}{\sum_{i=1}^{M-1} w_\theta(X^i,y,U^i)}\right] - \frac{w_\theta(x,y,u)}{\mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right]} \right) \nu(du)\,\tilde{\boldsymbol{\mu}}(dx) \right\|$$
$$\le \left| \frac{1}{\mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right]} - \mathbb{E}_{(\mu\otimes\nu)^{\otimes(M-1)}}\left[\frac{M}{\sum_{i=1}^{M-1} w_\theta(X^i,y,U^i)}\right] \right| \left\| \int w_\theta(x,y,u)\,\nu(du)\,\tilde{\boldsymbol{\mu}}(dx) \right\|$$
$$\le \frac{\tilde{b}_2 \epsilon^{-3}\,\|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}}}{M-1}. \quad (25)$$

Finally, we turn to the last term of (22) and proceed like

$$\left\| \int M \sum_{j=2}^\infty \frac{(-1)^{j+1}}{j} \mathbb{E}_{(\mu\otimes\nu)^{\otimes(M-1)}}\left[\left(\frac{w_\theta(x,y,u)}{\sum_{i=1}^{M-1} w_\theta(X^i,y,U^i)}\right)^j\right] \nu(du)\,\tilde{\boldsymbol{\mu}}(dx) \right\|$$
$$\le \int \left| M \sum_{j=2}^\infty \frac{(-1)^{j+1}}{j} \mathbb{E}_{(\mu\otimes\nu)^{\otimes(M-1)}}\left[\left(\frac{w_\theta(x,y,u)}{\sum_{i=1}^{M-1} w_\theta(X^i,y,U^i)}\right)^j\right] \right| \nu(du)\,|\tilde{\boldsymbol{\mu}}|\,(dx)$$
$$\le \|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}}\, M \sum_{j=2}^\infty \frac{1}{j}\left(\frac{\epsilon^{-6}}{M-1}\right)^j \le 2\epsilon^{-6}\,\|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}} \sum_{j=2}^\infty \frac{1}{j}\left(\frac{\epsilon^{-6}}{M-1}\right)^{j-1}. \quad (26)$$

Combining (24), (25), and (26) we obtain

$$\left\|G_\theta^M(z) - G_\theta(z)\right\| \le \left\| \mathbb{E}_{(\mu\otimes\nu)^{\otimes M}}\left[\frac{\sum_{i=1}^M \nabla_\theta w_\theta(X^i,y,U^i)}{\sum_{i=1}^M w_\theta(X^i,y,U^i)}\right] - \frac{\mathbb{E}_{\mu\otimes\nu}\left[\nabla_\theta w_\theta(X,y,U)\right]}{\mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right]} \right\|$$
$$+ \left\| \int \left(\mathbb{E}_{(\mu\otimes\nu)^{\otimes(M-1)}}\left[\frac{M w_\theta(x,y,u)}{\sum_{i=1}^{M-1} w_\theta(X^i,y,U^i)}\right] - \frac{w_\theta(x,y,u)}{\mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right]} \right) \nu(du)\,\tilde{\boldsymbol{\mu}}(dx) \right\|$$
$$+ \left\| \int M \sum_{j=2}^\infty \frac{(-1)^{j+1}}{j} \left(\mathbb{E}_{(\mu\otimes\nu)^{\otimes(M-1)}}\left[\frac{w_\theta(x,y,u)}{\sum_{i=1}^{M-1} w_\theta(X^i,y,U^i)}\right] \right)^j \nu(du)\,\tilde{\boldsymbol{\mu}}(dx) \right\|$$
$$\le \frac{\tilde{b}_1}{M} + \frac{(\tilde{b}_2\epsilon^{-3} + 2\epsilon^{-6})\,\|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}}}{M-1} + 2\epsilon^{-6}\,\|\tilde{\boldsymbol{\mu}}\|_{\mathsf{TV}} \sum_{j=2}^\infty \frac{1}{j+1}\left(\frac{\epsilon^{-6}}{M-1}\right)^j,$$

which concludes the proof. □

Corollary A.30. *Let Assumptions A.1, A.2, A.17, A.24, and A.18 hold. Then there exists a function $\beta : \mathbb{N}_{\ge 2} \to \mathbb{R}_+^*$ such that*

$$\beta(M) = \mathcal{O}\left(\frac{1}{M-1}\right),$$

and which satisfies, for all $M \geq 2$, $\theta \in \Theta$, $t \in \mathbb{N}$, and $y_{1:t+1} \in \mathsf{Y}^{t+1}$,

$$\left\|G_\theta^M(y_{t+1}, \phi_t^\theta, \psi_t^\theta) - G_\theta(y_{t+1}, \phi_t^\theta, \psi_t^\theta)\right\| \leq \beta(M).$$

Consequently, it also holds that

$$\left\|\nabla_\theta \ell(\theta) - \nabla_\theta \ell^M(\theta)\right\| \leq \beta(M).$$

Proof. By Theorem A.29, if $M > \epsilon^{-6} + 1$,

$$\left\|G_\theta^M(y_{t+1}, \phi_t^\theta, \psi_t^\theta) - G_\theta(y_{t+1}, \phi_t^\theta, \psi_t^\theta)\right\| \leq \frac{\tilde{b}_1}{M} + \frac{(\tilde{b}_2\epsilon^{-3} + 2\epsilon^{-6})\left\|\psi_t^\theta\right\|_{\mathsf{TV}} + 2\epsilon^{-6}\left\|\psi_t^\theta\right\|_{\mathsf{TV}}}{M-1} \sum_{j=2}^{\infty} \frac{1}{j+1}\left(\frac{\epsilon^{-6}}{M-1}\right)^j,$$

otherwise, inspecting the proof of Lemma A.26,

$$\left\|G_\theta^M(y_{t+1}, \phi_t^\theta, \psi_t^\theta) - G_\theta(y_{t+1}, \phi_t^\theta, \psi_t^\theta)\right\| \leq (2\tilde{\kappa}_2\epsilon^{-6} + 2\epsilon^{-2}\tilde{\kappa}_1)(1 + \left\|\psi_t^\theta\right\|_{\mathsf{TV}}).$$

Note that by Lemma A.15, $\left\|\psi_t^\theta\right\|_{\mathsf{TV}} \leq c_\psi$. Hence, we may define

$$\beta(M) := (\tilde{b}_1 + \tilde{b}_2 + 2)c_\psi \frac{\epsilon^{-6}}{M-1} + 2\epsilon^{-6}c_\psi \sum_{j=2}^{\infty} \frac{1}{j+1}\left(\frac{\epsilon^{-6}}{M-1}\right)^j \mathbb{1}_{\{M > \epsilon^{-6}+1\}} + (2\tilde{\kappa}_2\epsilon^{-6} + 2\epsilon^{-2}\tilde{\kappa}_1)(1 + c_\psi)\mathbb{1}_{\{M \leq \epsilon^{-6}+1\}},$$

which is clearly $\mathcal{O}(1/(M-1))$, proving that the first claim holds true. Finally, by applying Proposition A.27 we obtain that

$$\left\|\nabla_\theta \ell(\theta) - \nabla_\theta \ell^M(\theta)\right\| \leq \inf_{t \in \mathbb{N}_{>0}} \left\{ \left\| \nabla_\theta \ell(\theta) - \frac{1}{t}\sum_{s=0}^{t-1} \mathbb{E}\left[G_\theta(Z_s^\theta)\right] \right\| + \left\| \frac{1}{t}\sum_{s=0}^{t-1} \mathbb{E}\left[G_\theta^M(Z_s^\theta)\right] - \nabla_\theta \ell^M(\theta) \right\| + \frac{1}{t}\sum_{s=0}^{t-1} \mathbb{E}\left[\left\|G_\theta(Z_s^\theta) - G_\theta^M(Z_s^\theta)\right\|\right] \right\} \leq \beta(M),$$

which concludes the proof. □

A.6 Auxiliary results

Lemma A.31 (Convergence of martingale difference sequences)**.** *Let $(N_t)_{t \in \mathbb{N}_{>0}}$ be a martingale difference sequence in $\mathbb{R}$,* i.e. *an adapted stochastic process such that $\mathbb{E}[N_{t+1} \mid N_t, \ldots, N_0] = 0$ for all $t \in \mathbb{N}$, and let*

$$\sup_{t \in \mathbb{N}} \mathbb{E}[N_t^2] < c,$$

for some finite constant $c > 0$. Then

$$\lim_{T \to \infty} \frac{1}{T}\sum_{t=1}^{T} N_t = 0, \quad \mathbb{P}\text{-a.s.}$$

Proof. For every $t \in \mathbb{N}_{>0}$, let $M_t := \sum_{s=1}^{t} N_s/s$. It is easy to check that $(M_t)_{t \in \mathbb{N}_{>0}}$ is a martingale. Moreover, for all t we have

$$\mathbb{E}[M_t^2] = \sum_{s=1}^{t}\sum_{s'=1}^{t} \mathbb{E}\left[\frac{N_s N_{s'}}{ss'}\right] = \sum_{s=1}^{t} \mathbb{E}\left[\frac{N_s^2}{s^2}\right] + 2\sum_{s=1}^{t}\sum_{s'=1}^{s-1} \mathbb{E}\left[\frac{N_{s'}\mathbb{E}[N_s \mid N_{s-1}, \ldots, N_0]}{ss'}\right] \leq c\sum_{s=1}^{t} \mathbb{E}\left[\frac{1}{s^2}\right] < 2c.$$

Then by the *martingale convergence theorem* (see, *e.g.*, Gut, 2013, Theorem 12.2, page 518) it has a limit $\mathbb{P}$-almost surely. Finally, by Kronecker's lemma, since $\sum_{t=1}^{T} N_t/t$ converges $\mathbb{P}$-a.s. as $T \to \infty$, then

$$\lim_{T \to \infty} \frac{1}{T}\sum_{t=1}^{T} N_t = 0, \quad \mathbb{P}\text{-a.s.}$$

□

Lemma A.32. *Let Assumptions A.1–A.2, and A.24 hold and let $M > \epsilon^{-6} + 1$. Then for all $z = (x, y, \mu, \tilde{\boldsymbol{\mu}}) \in \mathsf{Z}$ it holds*

$$\int M \log\left(1 + \frac{w_\theta(x,y,u)}{\sum_{i=1}^{M-1} w_\theta(x^i,y,u^i)}\right) \nu(du)\,\tilde{\boldsymbol{\mu}}(dx) = \int M \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} \left(\frac{w_\theta(x,y,u)}{\sum_{i=1}^{M-1} w_\theta(x^i,y,u^i)}\right)^k \nu(du)\,\tilde{\boldsymbol{\mu}}(dx).$$

Proof. The claim follows immediately from a Taylor expansion, $\log(1+x)$ being analytic when $|x| < 1$ and

$$\frac{w_\theta(x,y,u)}{\sum_{i=1}^{M-1} w_\theta(x^i,y,u^i)} \leq \frac{\epsilon^{-6}}{M-1} < 1$$

for $M > \epsilon^{-6} + 1$. □

Lemma A.33. *Let Assumptions A.1, A.2, and A.24 hold. Then for every $\mu \in \mathsf{M}_1(\mathcal{X})$, $y \in \mathsf{Y}$, and $M \geq 2$ there exist constants $\tilde{b}_1 > 0$ and $\tilde{b}_2 > 0$, depending only on ϵ and $\tilde{\kappa}_2$, such that*

$$\left\| \mathbb{E}_{(\mu\otimes\nu)^{\otimes M}} \left[\frac{\sum_{i=1}^M \nabla_\theta w_\theta(X^i,y,U^i)}{\sum_{i=1}^M w_\theta(X^i,y,U^i)} \right] - \frac{\mathbb{E}_{\mu\otimes\nu}\left[\nabla_\theta w_\theta(X,y,U)\right]}{\mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right]} \right\| \leq \frac{\tilde{b}_1}{M}$$

and

$$\left| \frac{1}{\mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right]} - \mathbb{E}_{(\mu\otimes\nu)^{\otimes(M-1)}} \left[\frac{M}{\sum_{i=1}^{M-1} w_\theta(X^i,y,U^i)} \right] \right| \leq \frac{\tilde{b}_2}{M-1}.$$

Proof. We apply the identity $a/b - c/d = 1/d(a/b(d-b) + (a-c))$ and write

$$\begin{aligned}
&\frac{\mathbb{E}_{\mu\otimes\nu}\left[\nabla_\theta w_\theta(X,y,U)\right]}{\mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right]} - \mathbb{E}_{(\mu\otimes\nu)^{\otimes M}} \left[\frac{\sum_{i=1}^M \nabla_\theta w_\theta(X^i,y,U^i)/M}{\sum_{i=1}^M w_\theta(X^i,y,U^i)/M} \right] \\
&\quad = \mathbb{E}_{(\mu\otimes\nu)^{\otimes M}} \left[\frac{1}{\sum_{i=1}^M w_\theta(X^i,y,U^i)/M} \left(\frac{\mathbb{E}_{\mu\otimes\nu}\left[\nabla_\theta w_\theta(X,y,U)\right]}{\mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right]} \right.\right. \\
&\left.\left. \left(\frac{1}{M}\sum_{i=1}^M w_\theta(X^i,y,U^i) - \mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right] \right) + \left(\mathbb{E}_{\mu\otimes\nu}\left[\nabla_\theta w_\theta(X,y,U)\right] - \frac{1}{M}\sum_{i=1}^M \nabla_\theta w_\theta(X^i,y,U^i) \right) \right) \right] \\
&\quad = \mathbb{E}_{(\mu\otimes\nu)^{\otimes M}} \left[\left(\frac{1}{\sum_{i=1}^M w_\theta(X^i,y,U^i)/M} - \frac{1}{\mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right]} \right) \right. \\
&\quad\quad \times \left(\frac{\mathbb{E}_{\mu\otimes\nu}\left[\nabla_\theta w_\theta(X,y,U)\right]}{\mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right]} \frac{1}{M}\sum_{i=1}^M \left(w_\theta(X^i,y,U^i) - \mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right] \right) \right. \\
&\quad\quad\quad \left.\left. + \frac{1}{M}\sum_{i=1}^M \left(\mathbb{E}_{\mu\otimes\nu}\left[\nabla_\theta w_\theta(X,y,U)\right] - \nabla_\theta w_\theta(X^i,y,U^i) \right) \right) \right],
\end{aligned}$$

where we used $\mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U) - \mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right]\right] = 0$ and $\mathbb{E}_{\mu\otimes\nu}\left[\nabla_\theta w_\theta(X,y,U) - \mathbb{E}_{\mu\otimes\nu}\left[\nabla_\theta w_\theta(X,y,U)\right]\right] = 0$ in the last equality. Using Lemma A.25 and the Cauchy–Schwarz inequality,

$$\begin{aligned}
&\left\| \frac{\mathbb{E}_{\mu\otimes\nu}\left[\nabla_\theta w_\theta(X,y,U)\right]}{\mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right]} - \mathbb{E}_{(\mu\otimes\nu)^{\otimes M}} \left[\frac{\sum_{i=1}^M \nabla_\theta w_\theta(X^i,y,U^i)/M}{\sum_{i=1}^M w_\theta(X^i,y,U^i)/M} \right] \right\| \\
&\quad \leq \mathbb{E}_{(\mu\otimes\nu)^{\otimes M}} \left[\frac{\left| \frac{1}{M}\sum_{i=1}^M w_\theta(X^i,y,U^i) - \mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right] \right|}{\mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right] \sum_{i=1}^M w_\theta(X^i,y,U^i)/M} \right. \\
&\quad\quad \left. \times \frac{\mathbb{E}_{\mu\otimes\nu}\left[\|\nabla_\theta w_\theta(X,y,U)\|\right]}{\mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right]} \left| \frac{1}{M}\sum_{i=1}^M w_\theta(X^i,y,U^i) - \mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right] \right| \right]
\end{aligned}$$

$$
\begin{aligned}
&+ \mathbb{E}_{(\mu\otimes\nu)^{\otimes M}}\left[\frac{\left|\frac{1}{M}\sum_{i=1}^{M} w_\theta(X^i,y,U^i) - \mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right]\right|}{\mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right]\sum_{i=1}^{M} w_\theta(X^i,y,U^i)/M}\right.\\
&\left.\times\left\|\frac{1}{M}\sum_{i=1}^{M}\nabla_\theta w_\theta(X^i,y,U^i) - \mathbb{E}_{\mu\otimes\nu}\left[\nabla_\theta w_\theta(X,y,U)\right]\right\|\right]\\
&\leq \epsilon^{-9}\tilde{\kappa}_2\mathbb{E}_{(\mu\otimes\nu)^{\otimes M}}\left[\left(\frac{1}{M}\sum_{i=1}^{M} w_\theta(X^i,y,U^i) - \mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right]\right)^2\right]\\
&\quad+\epsilon^{-6}\mathbb{E}_{(\mu\otimes\nu)^{\otimes M}}\left[\left(\frac{1}{M}\sum_{i=1}^{M} w_\theta(X^i,y,U^i) - \mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right]\right)^2\right]^{1/2}\\
&\quad\times\mathbb{E}_{(\mu\otimes\nu)^{\otimes M}}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\nabla_\theta w_\theta(X^i,y,U^i) - \mathbb{E}_{\mu\otimes\nu}\left[\nabla_\theta w_\theta(X,y,U)\right]\right\|^2\right]^{1/2}\\
&\leq \epsilon^{-9}\tilde{\kappa}_2\frac{1}{M}\mathbb{V}_{\mu\otimes\nu}(w_\theta(X,y,U)) + \epsilon^{-6}\frac{1}{M}\mathbb{V}_{\mu\otimes\nu}(w_\theta(X,y,U))^{1/2}\\
&\quad\times\mathbb{E}_{\mu\otimes\nu}\left[\left\|\nabla_\theta w_\theta(X,y,U) - \mathbb{E}_{\mu\otimes\nu}\left[\nabla_\theta w_\theta(X,y,U)\right]\right\|^2\right]^{1/2} \leq 3\epsilon^{-9}\tilde{\kappa}_2\frac{1}{M} =: \frac{\tilde{b}_1}{M}.
\end{aligned}
$$

Similarly, to establish the second claim of the lemma we proceed like

$$
\begin{aligned}
&\left|\frac{1}{\mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right]} - \mathbb{E}_{(\mu\otimes\nu)^{\otimes(M-1)}}\left[\frac{M/(M-1)}{\sum_{i=1}^{M-1} w_\theta(X^i,y,U^i)/(M-1)}\right]\right|\\
&\quad= \mathbb{E}_{(\mu\otimes\nu)^{\otimes(M-1)}}\left[\left(\frac{1}{\sum_{i=1}^{M-1} w_\theta(X^i,y,U^i)/(M-1)} - \frac{1}{\mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right]}\right)\right.\\
&\qquad\left.\times\frac{1}{\mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right]}\left(\frac{1}{M-1}\sum_{i=1}^{M-1} w_\theta(X^i,y,U^i) - \mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right]\right)\right]\\
&\qquad+\mathbb{E}_{(\mu\otimes\nu)^{\otimes(M-1)}}\left[\frac{1-M/(M-1)}{\sum_{i=1}^{M-1} w_\theta(X^i,y,U^i)/(M-1)}\right]\\
&\quad\leq \mathbb{E}_{(\mu\otimes\nu)^{\otimes(M-1)}}\left[\frac{\left|\mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right] - \sum_{i=1}^{M-1} w_\theta(X^i,y,U^i)/(M-1)\right|}{\mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right]\sum_{i=1}^{M-1} w_\theta(X^i,y,U^i)/(M-1)}\right.\\
&\qquad\left.\times\frac{1}{\mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right]}\left|\frac{1}{M-1}\sum_{i=1}^{M-1} w_\theta(X^i,y,U^i) - \mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right]\right|\right]\\
&\qquad+\mathbb{E}_{(\mu\otimes\nu)^{\otimes(M-1)}}\left[\frac{1/(M-1)}{\sum_{i=1}^{M-1} w_\theta(X^i,y,U^i)/(M-1)}\right]\\
&\quad\leq \epsilon^{-9}\mathbb{E}_{(\mu\otimes\nu)^{\otimes(M-1)}}\left[\left(\frac{1}{M-1}\sum_{i=1}^{M-1} w_\theta(X^i,y,U^i) - \mathbb{E}_{\mu\otimes\nu}\left[w_\theta(X,y,U)\right]\right)^2\right]\\
&\qquad+\epsilon^{-3}\frac{1}{M-1} \leq (\epsilon^{-15}+\epsilon^{-3})\frac{1}{M-1} =: \frac{\tilde{b}_2}{M-1}.
\end{aligned}
$$

□

B Proof of Lemma 4.1

Proof. First note that the complete-data score function is additive, *i.e.* it satisfies

$$\nabla_\theta \log p_\theta(x_{0:t}, y_{0:t}) = \nabla_\theta \log g_\theta(y_0 \mid x_0) + \sum_{s=0}^{t-1} \nabla_\theta \log m_\theta(x_{s+1} \mid x_s) + \nabla_\theta \log g_\theta(y_{s+1} \mid x_{s+1})$$

for all $t \in \mathbb{N}$, $x_{0:t} \in \mathsf{X}^{t+1}$ and $y_{0:t} \in \mathsf{Y}^{t+1}$. Now, for $t \in \mathbb{N}_{>0}$ we write

$$\begin{aligned}
\varphi^\theta_{t+1}(x_{t+1}) &= \int \nabla_\theta \log p_\theta(x_{0:t+1}, y_{0:t+1}) p_\theta(x_{0:t} \mid y_{0:t}, x_{t+1})\, dx_{0:t} \\
&= \int \nabla_\theta \log p_\theta(x_{0:t+1}, y_{0:t+1}) p_\theta(x_{0:t-1} \mid y_{0:t}, x_{t:t+1}) p_\theta(x_t \mid y_{0:t}, x_{t+1})\, dx_{0:t} \\
&= \iint \nabla_\theta \log p_\theta(x_{0:t}, y_{0:t}) p_\theta(x_{0:t-1} \mid y_{0:t-1}, x_t)\, dx_{0:t-1}\, p_\theta(x_t \mid y_{0:t}, x_{t+1})\, dx_t \\
&\quad + \int \nabla_\theta \log\{m_\theta(x_{t+1} \mid x_t) g_\theta(y_{t+1} \mid x_{t+1})\} \int p_\theta(x_{0:t-1} \mid y_{0:t-1}, x_t)\, dx_{0:t-1}\, p_\theta(x_t \mid y_{0:t}, x_{t+1})\, dx_t \\
&= \int \left(\varphi^\theta_t(x_t) + \nabla_\theta \log\{m_\theta(x_{t+1} \mid x_t) g_\theta(y_{t+1} \mid x_{t+1})\}\right) \frac{\phi^\theta_t(x_t) m_\theta(x_{t+1} \mid x_t)}{\int \phi^\theta_t(x) m_\theta(x_{t+1} \mid x)\, dx}\, dx_t.
\end{aligned}$$

Hence, given $(\varphi^\theta_t, \phi^\theta_t, y_{t+1})$, we define the recursive update

$$\bar\Psi_\theta(\varphi^\theta_t, \phi^\theta_t, y_{t+1})(x_{t+1}) := \int \left(\varphi^\theta_t(x_t) + \nabla_\theta \log\{m_\theta(x_{t+1} \mid x_t) g_\theta(y_{t+1} \mid x_{t+1})\}\right) \frac{\phi^\theta_t(x_t) m_\theta(x_{t+1} \mid x_t)}{\int \phi^\theta_t(x) m_\theta(x_{t+1} \mid x)\, dx}\, dx_t,$$

which satisfies $\varphi^\theta_{t+1}(x_{t+1}) = \bar\Psi_\theta(\varphi^\theta_t, \phi^\theta_t, y_{t+1})(x_{t+1})$ for all $x_{t+1} \in \mathsf{X}$. This proves (i).

To prove (ii), we first note that for every $t \in \mathbb{N}$ and every measurable function $f : \mathsf{X} \to \mathbb{R}$,

$$\begin{aligned}
\int f(x_t) \psi^\theta_t(x_t)\, dx_t &= \int f(x_t) \nabla_\theta \phi^\theta_t(x_t)\, dx_t = \int f(x_t) \nabla_\theta \frac{p_\theta(x_{0:t}, y_{0:t})}{p_\theta(y_{0:t})}\, dx_{0:t} \\
&= \int f(x_t) \nabla_\theta \log p_\theta(x_{0:t}, y_{0:t}) p_\theta(x_{0:t} \mid y_{0:t})\, dx_{0:t} - \frac{\nabla_\theta p_\theta(y_{0:t})}{p_\theta(y_{0:t})} \int f(x_t) \phi^\theta_t(x_t)\, dx_t \\
&= \mathbb{E}_{\phi^\theta_{0:t}}[f(X_t)(\nabla_\theta \log p_\theta(X_{0:t}, y_{0:t}) - \mathbb{E}_{\phi^\theta_{0:t}}[\nabla_\theta \log p_\theta(X_{0:t}, y_{0:t})])],
\end{aligned}$$

which implies

$$\begin{aligned}
G^M_\theta(y_{t+1}, \phi^\theta_t, \psi^\theta_t) = \mathbb{E}_{(\phi^\theta_t)^{\otimes M-1} \otimes \phi^\theta_{0:t} \otimes \nu^{\otimes M}} \Bigg[&\frac{\sum_{i=1}^M \nabla_\theta w_\theta(X^i, y_{t+1}, U^i)}{\sum_{i'=1}^M w_\theta(X^{i'}, y_{t+1}, U^{i'})} \\
&+ M \log\left(\frac{1}{M} \sum_{i=1}^M w_\theta(X^i, y_{t+1}, U^i)\right) \left(\nabla_\theta \log p_\theta(X^M_{0:t}, y_{0:t}) - \mathbb{E}_{\phi^\theta_{0:t}}[\nabla_\theta \log p_\theta(X_{0:t}, y_{0:t})]\right) \Bigg],
\end{aligned}$$

where $(X^i)_{i=1}^{M-1}$ are i.i.d. draws from ϕ^θ_t, the trajectory $X^M_{0:t}$ is drawn from $\phi^\theta_{0:t}$, and the auxiliary variables $(U^i)_{i=1}^M$ are i.i.d. draws from ν. Now, by the definition of φ^θ_t, we have

$$\begin{aligned}
\mathbb{E}_{\phi^\theta_{0:t}}[f(X_t) \nabla_\theta \log p_\theta(X_{0:t}, y_{0:t})] &= \int f(x_t) \nabla_\theta \log p_\theta(x_{0:t}, y_{0:t}) \phi^\theta_{0:t}(x_{0:t})\, dx_{0:t} \\
&= \int f(x_t) \nabla_\theta \log p_\theta(x_{0:t}, y_{0:t}) p_\theta(x_{0:t-1} \mid y_{0:t}, x_t) p_\theta(x_t \mid y_{0:t})\, dx_{0:t} \\
&= \int f(x_t) \int \nabla_\theta \log p_\theta(x_{0:t}, y_{0:t}) p_\theta(x_{0:t-1} \mid y_{0:t-1}, x_t)\, dx_{0:t-1}\, \phi^\theta_t(x_t)\, dx_t \\
&= \mathbb{E}_{\phi^\theta_t}[f(X_t) \varphi^\theta_t(X_t)].
\end{aligned}$$

Thus, it finally holds that

$$G_\theta^M(y_{t+1}, \phi_t^\theta, \psi_t^\theta) = \mathbb{E}_{(\phi_t^\theta \otimes \nu)^{\otimes M}} \left[\frac{\sum_{i=1}^M \nabla_\theta w_\theta(X^i, y_{t+1}, U^i)}{\sum_{i'=1}^M w_\theta(X^{i'}, y_{t+1}, U^{i'})} + M \log \left(\frac{1}{M} \sum_{i=1}^M w_\theta(X^i, y_{t+1}, U^i) \right) \left(\varphi_t^\theta(X^M) - \mathbb{E}_{\phi_t^\theta}[\varphi_t^\theta(X)] \right) \right] = \bar{G}_\theta^M(y_{t+1}, \phi_t^\theta, \varphi_t^\theta).$$

□

C The `AdaSmooth` algorithm

In this appendix we discuss how to produce recursively by means of SMC methods (Gordon et al., 1993; Doucet et al., 2001; Chopin and Papaspiliopoulos, 2020) the sample $\{(\xi_{t+1}^i, \tau_{t+1}^i, \omega_{t+1}^i)\}_{i=1}^N$, given $\{(\xi_t^i, \tau_t^i, \omega_t^i)\}_{i=1}^N$ together with the new observation Y_{t+1} and the updated parameter θ_{t+1}. In a standard particle filter, the propagation of the particles is carried out by first resampling, with replacement, the particles $(\xi_t^i)_{i=1}^N$ proportionally to the weights $(\omega_t^i)_{i=1}^N$ (the so-called *selection* step) and then *mutating* the resampled particles using some proposal distribution. Here we let the proposal be r_θ, which is progressively adapted over the iterations, but also other choices are possible. Then, after the recalculation of the importance weights, it remains to propagate the terms $(\tau_t^i)_{i=1}^N$. Several works on particle-based online additive smoothing have presented strategies for updating these statistics by approximating the recursion $\bar{\Psi}_\theta$. In fact, this requires the computation of an expectation with respect to the so-called *backward kernel*, whose density is proportional to $\phi_t^\theta(x_t) m_\theta(x_{t+1} \mid x_t)$; see Appendix B for details. For each propagated particle ξ_{t+1}^i, $i \in \{1, \dots, N\}$, the backward kernel is translated into a categorical distribution with support on the previous particle cloud $(\xi_t^j)_{j=1}^N$ and probabilities proportional to $\{\omega_t^j m_\theta(\xi_{t+1}^i \mid \xi_t^j)\}_{j=1}^N$. Del Moral et al. (2010) evaluate these expectations exactly, while the `PaRIS` algorithm (Olsson and Westerborn, 2017) employs a Monte Carlo approximation based on two samples at least (sufficient to guarantee long-term stability) from $\mathsf{cat}(\{\omega_t^j m_\theta(\xi_{t+1}^i \mid \xi_t^j)\}_{j=1}^N)$, for each ξ_{t+1}^i, $i \in \{1, \dots, N\}$. Both of these approaches have complexity $\mathcal{O}(N^2)$ per time step, due to the computation of the normalising constants, which is not desirable; therefore, the latter implements the backward sampling according to the accept-reject alternative suggested by Douc et al. (2011), where the complexity of each draw does not depend on N. Still, Dau and Chopin (2023), showed that in many realistic cases, the expected time to acceptance is infinite, which forces to have an early stopping rule for the rejection sampler and obtain the remaining draws from the exact distributions. Therefore, since backward sampling remains the bottleneck of this class of methods, the `AdaSmooth` algorithm (Mastrototaro et al., 2024) proposes to reduce the frequency of this operation by combining a fast but unstable (as the asymptotic variance grows quadratically in time rather than linearly) naive forward smoother with the `PaRIS`. This is accomplished by applying sparsely the resampling and the backward sampling operations according to a schedule which is governed by two sequences $(\rho_t^{\mathsf{res}})_{t \in \mathbb{N}} \in \{0, 1\}$ and $(\rho_t^{\mathsf{bs}})_{t \in \mathbb{N}} \in \{0, 1\}$, indicating whether to perform resampling or backward sampling at each iteration, respectively. These sequences may be deterministic or adapted to the random variables generated by the algorithm, *e.g.*, by monitoring the weights and the particle-path degeneracy, respectively. If the values equal to one appear regularly in the sequences, then the algorithm is proved to be stable (see Mastrototaro et al., 2024, for more details). Algorithm C.1 illustrates the `AdaSmooth` update in the special case where the additive functional is the complete-data score of the SSM.

D Additional numerical experiments

In this appendix, we present additional results from the numerical experiments conducted on the SLAM model to further highlight the advantages of the `OSIWAE` algorithm over `RML` and `OVSMC`. Figure 7 illustrates the Mean Absolute Error (MAE) averaged over 10 runs for the estimated positions of all landmarks. The dashed lines represent the minimum and maximum MAE observed across all runs at each timestep. As discussed in the main paper, we observe that after the initial phase where the proposal distribution is being learned, `OSIWAE` achieves a lower MAE than both other algorithms.

Algorithm C.1 `AdaSmooth`

Require: $\{(\xi_t^i, \tau_t^i, \omega_t^i)\}_{i=1}^N, Y_{t+1}, \theta_{t+1}$.

for $i \leftarrow 1, \dots, N$ **do**
  if ρ_t^{res} **then**
    draw $I_{t+1}^i \sim \mathsf{cat}(\{\omega_t^j\}_{j=1}^N)$;
  else
    set $I_{t+1}^i \leftarrow i$;
  end if
  draw $\xi_{t+1}^i \sim r_{\theta_{t+1}}(\cdot \mid \xi_t^{I_{t+1}^i}, Y_{t+1})$;
  set $\omega_{t+1}^i \leftarrow (\omega_t^i)^{1-\rho_t^{\text{res}}} \dfrac{m_{\theta_{t+1}}(\xi_{t+1}^i \mid \xi_t^{I_{t+1}^i}) g_{\theta_{t+1}}(Y_{t+1} \mid \xi_{t+1}^i)}{r_{\theta_{t+1}}(\xi_{t+1}^i \mid \xi_t^{I_{t+1}^i}, Y_{t+1})}$;
  if ρ_t^{res} **and** ρ_t^{bs} **then**
    draw $J_{t+1}^i \sim \mathsf{cat}(\{\omega_t^j m_{\theta_{t+1}}(\xi_{t+1}^i \mid \xi_t^j)\}_{j=1}^N)$;
    set $\tau_{t+1}^i \leftarrow \dfrac{1}{2}\big(\tau_t^{I_{t+1}^i} + \nabla_\theta \log m_{\theta_{t+1}}(\xi_{t+1}^i \mid \xi_t^{I_{t+1}^i}) + \tau_t^{J_{t+1}^i} + \nabla_\theta \log m_{\theta_{t+1}}(\xi_{t+1}^i \mid \xi_t^{J_{t+1}^i})\big)$
      $+ \nabla_\theta \log g_{\theta_{t+1}}(Y_{t+1} \mid \xi_{t+1}^i)$;
  else
    set $\tau_{t+1}^i \leftarrow \tau_t^{I_{t+1}^i} + \nabla_\theta \log m_{\theta_{t+1}}(\xi_{t+1}^i \mid \xi_t^{I_{t+1}^i}) + \nabla_\theta \log g_{\theta_{t+1}}(Y_{t+1} \mid \xi_{t+1}^i)$;
  end if
end for
return $\{(\xi_{t+1}^i, \tau_{t+1}^i, \omega_{t+1}^i)\}_{i=1}^N$.

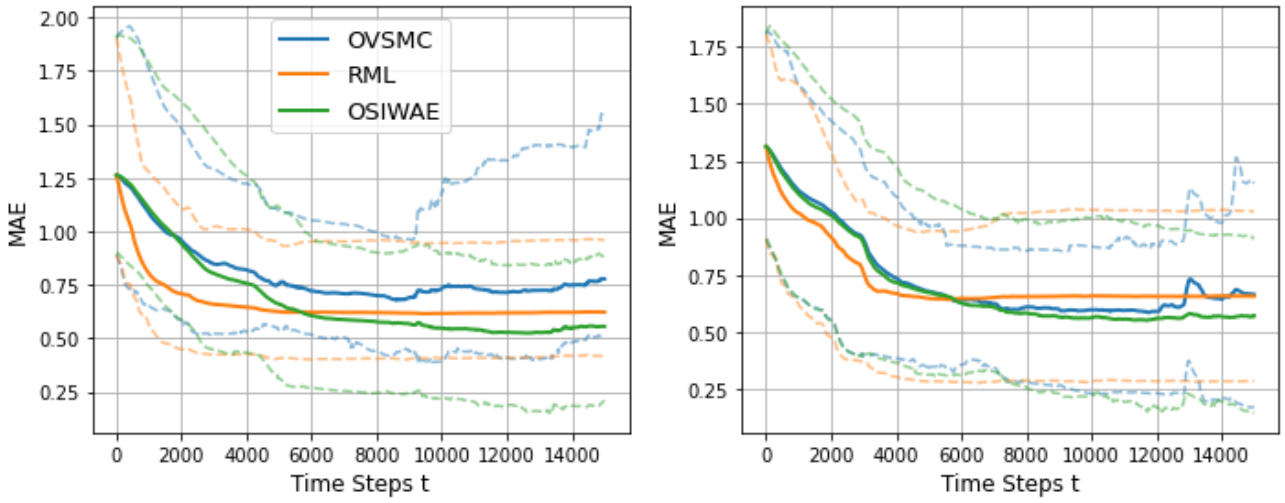

Figure 7: Mean absolute errors (MAEs) averaged over 10 runs for the estimated positions of $L = 8$ landmarks over time using `OSIWAE`, `RML`, and `OVSMC` in a SLAM scenario with motion noise variance $\sigma^2_{\text{motion}} = 0.2$ and observation noise variance $\sigma^2_{\text{obs}} = 0.1$. The dashed lines indicate the minimum and maximum MAE across all runs. The proposal distribution $r_\theta(\cdot \mid x_t, y_{t+1})$ in both `OSIWAE` and `OVSMC` is learned using two distinct neural networks, each with one hidden layer of 12 nodes. All three methods employ $N = 1000$ particles, and `OSIWAE` uses $M = 1000$ samples. Left panel: All three algorithms are run on the same data without any prior learning. Right panel: A training run is first performed using `SMC-OSIWAE` on a different dataset to learn the proposal distribution; subsequently, all three algorithms are applied to the same data. Each of the 10 runs is performed with the same observations and true landmark positions but with different initial landmark estimates.