

## A Appendix

The appendix contains proofs and implementation details for the main paper. It is organized as follows:

1. Related work Appendix [B](#)
2. Background Appendix [C](#)
  - Exponential family distributions Appendix [C.1](#)
  - Mean field variational inference Appendix [C.2](#)
  - Bayes estimation Appendix [C.3](#)
3. Message passing [D](#)
  - Sequential message passing [D.1](#)
  - Parallel message passing [D.2](#)
  - Basic probabilistic queries [D.4](#)
4. Conditioned linear SDEs [E](#)
  - Conditioned linear SDEs [E.1](#)
  - Basic probabilistic queries [E.2](#)
  - Corresponding probability flow ODE [E.3](#)
5. Constrained mean field VI [F](#)
  - Derivation [F.1](#)
  - Bayes estimator equivariance [F.2](#)
  - CMFVI time series models [F.3](#)
6. Flow-based generative models [G](#)
  - Score function of FBGMs [G.1](#)
  - General form of Markovian projection SDE [G.2](#)
  - General form of Markovian projection ODE [G.3](#)
7. Message passing implementation details [H](#)
  - Numerical stability considerations [H.1](#)
  - Message passing pseudocode [H.2](#)
8. Dataset details [I](#)
9. Model implementation details [J](#)

## B Related Work

There are numerous perspectives on flow-based generative models [\[Luo, 2022, Dieleman, 2023\]](#) and even more variants of these models. At their core, these models start by constructing a stochastic process that starts at a prior distribution and ends at the data distribution. Diffusion models use progressive noising of data to build this map [\[Sohl-Dickstein et al., 2015, Ho et al., 2020, Song et al., 2021\]](#) via a simple SDE whose stationary distribution is Gaussian. On the other hand, flow-matching models [\[Liu et al., 2023, Albergo and Vanden-Eijnden, 2023, Lipman et al., 2023\]](#) use a stochastic bridge to build this map by conditioning a simple SDE to start at a point in the prior distribution and end at the data distribution. The choice of simple SDE used in all of these models is a user-defined choice that typically is a linear SDE, such as variance preserving SDE [\[Song et al., 2021\]](#), Brownian motion, Ornstein-Uhlenbeck process, and others, due to their tractability as Gaussian processes [\[Särkkä and Solin, 2019\]](#), and is even used to construct more exotic latent SDEs such as critically damped Langevin dynamics [\[Dockhorn et al., 2022, Chen et al., 2024c\]](#) or the Weiner velocity model [\[Bar-Shalom et al., 2001, Särkkä et al., 2006\]](#). In our paper, we abstract away these choices and generally consider using linear SDEs to construct the initial map between distributions. There are a few different ways to go from this initial stochastic process to a FBGM. A common way to construct a FBGM from this is construct and optimize and ELBO for the likelihood of data under this initial process [\[Kingma et al., 2021\]](#). Alternatively, one can directly solve for the SDE whose marginal distribution is that of this initial process [\[Song et al., 2021, Lipman et al., 2023\]](#) or define it as the

SDE whose path measure is as close as possible to the initial process [Shi et al., 2024, De Bortoli et al., 2023] in terms of KL divergence, called the Markovian projection. We adopt the latter view over the ELBO view because it explicitly constructs a solution to the generative modeling problem and is available in closed form while this is hidden in the ELBO formulation and show that the solution to a mean field variational inference problem can be seen as an approximate discrete time counterpart.

Flow-based generative models have been successfully applied to time series problems in a *non-autoregressive* fashion [Kollovich et al., 2023, Yuan and Qiao, 2024, Kollovich et al., 2025, Hu et al., 2024, Yang et al., 2024, Meijer and Chen, 2024]. These models transform the time series generative modeling problem into the standard generative modeling problem used in image generation by treating each time series as a single vector by concatenating all times together, and then learning a map from a Gaussian vector of the same size to the data vector. These approaches can be conditioned using guidance [Rasul et al., 2021, Dhariwal and Nichol, 2021, Ho and Salimans, 2022, Kollovich et al., 2023] which allows them to perform tasks such as forecasting and imputation. Our approach differs from these in that we construct autoregressive models.

The class of models most relevant to our paper are autoregressive neural SDEs that are trained using principles from flow-based generative models. [Chen et al., 2024a] uses a Föllmer process to model the transition distributions of the distribution of time series data, which is the same approach that we adopt in our Neural SDE model. [Park et al., 2024] also learns a similar latent Neural SDE model that uses a similar form of soft conditioning as us (through the use of emission potentials), and is trained to maximize the likelihood of data. [Tamir et al., 2024] is also similar where they perform stochastic interpolation using Gaussian processes and perform inference with Kalman smoothing as well, which is a form of message passing. Finally, [Shen and Cheng, 2025] learns a more general SDE to learn the distribution of time series data where the diffusion coefficient is not independent of the current state and also maximize the likelihood of data. These related papers are all related to the Neural SDE that we describe in our paper. Our main contributions are centered around investigating how to apply the approach used to construct these continuous time models for creating similar discrete time models. [El-Gazzar and van Gerven, 2025] used flow matching to learn the next state distribution of time series data, but did not learn a Föllmer process for this task and instead learned to transform a Gaussian into the next state distribution.

## C Background

### C.1 Exponential family distributions

Our findings can be most easily written using exponential family distributions. Although we restrict our attention to Gaussian distributions, the form of our results are most readable in natural parameter space.

**Definition 3** (Exponential family distribution). *An probability distribution is in the exponential family if its density function can be written in the following form:*

$$p(x|\theta) = \exp\{\langle t(x), \theta \rangle - A(\theta)\} \quad (20)$$

where  $t(x)$  is called the sufficient statistic,  $\theta$  the natural parameter and  $A(\theta)$  the partition function.

The member of this family that we will use is the multivariate Gaussian distribution. A multivariate Gaussian with mean  $\mu$  and covariance matrix  $\Sigma$  has the sufficient statistic  $t(x) = (x, xx^T)$  and natural parameters  $\theta = (-\frac{1}{2}\Sigma^{-1}, \Sigma^{-1}\mu)$ . In practice, it is more convenient to drop the  $-\frac{1}{2}$  scaling term and work with the parameters  $(J, h) = (-\Sigma^{-1}, \Sigma^{-1}\mu)$ , where  $J$  is the precision matrix of the distribution. While these are not exactly the natural parameters, we will refer to them as so. Throughout this paper, we will work with unnormalized Gaussian distributions, which we call “Gaussian potentials”. We use the notation  $\phi(x|\theta)$  to denote a Gaussian potential function over  $x$  with natural parameters  $\theta$ . A convenient property of the natural parameter form is that the score function takes a simple form.

$$\nabla \log \phi(x|\theta) = Jx - h \quad (21)$$

Another Gaussian distribution that we will use extensively is the Gaussian transition distribution. We write  $\phi_{k+1|k}(x_{k+1}|x_k) = N(x_{k+1}|Ax_k + u, \Sigma)$  to denote the Gaussian transition distribution from  $x_k$  to  $x_{k+1}$  with state transition matrix  $A$ , bias vector  $u$  and covariance matrix  $\Sigma$ .

## 596 C.2 Mean field variational inference

597 Mean field variational inference is an approximate inference algorithm for probabilistic models. It's  
 598 main feature is that it's solution is available in a simple closed form expression. Let  $p(x, \theta)$  be a joint  
 599 distribution over  $x$  and  $\theta$ . The mean field variational problem is to find distributions,  $q_x(x)$  and  $q_\theta(\theta)$   
 600 that minimize the KL divergence between  $q_x(x)q_\theta(\theta)$  and  $p(x, \theta)$ .

601 **Proposition 8** (Mean field variational inference for CRFs). *Let  $p(\theta)$  be a distribution over  $\theta$ ,  $p(x|\theta)$   
 602 be the CRF in Definition 7 and  $p(x, \theta) = p(\theta)p(x|\theta)$  be the joint distribution over  $x$  and  $\theta$ . Then the  
 603 solutions to*

$$\operatorname{argmin}_{q_x(x), q_\theta(\theta)} \text{KL} [q_x(x)q_\theta(\theta) | p(x, \theta)] \quad (22)$$

604 will satisfy:

$$q_x(x) \propto \exp\{\mathbb{E}_{q_\theta(\theta)} [\log p(x|\theta)]\} \quad (23)$$

$$q_\theta(\theta) \propto \exp\{\mathbb{E}_{q_x(x)} [\log p(\theta|x)]\} \quad (24)$$

605 See [Beal, 2003] for a proof. Typical use cases of mean field VI use tractable classes of distributions  
 606 for  $p(\theta)$  and  $p(x|\theta)$  so that one can perform EM style, alternating updates to obtain the optimal  $q$   
 607 distributions [Beal, 2003, Johnson et al., 2014]. However, in our setting, we will use mean field VI  
 608 differently. We will assume nothing about the form of  $p(\theta)$ , but will constrain the variational problem  
 609 by fixing  $q_\theta(\theta) = p(\theta)$ .

## 610 C.3 Bayes estimation

611 **Lemma 1** (Bayes estimate of parameter). *Let  $p(z, \theta)$  be a joint distribution and let  $\theta^*(z)$  be the  
 612 Bayes estimate of  $\theta$  based on  $z$  under the squared error risk. Then the Bayes estimate takes the  
 613 following two forms:*

$$\theta^*(z) = \mathbb{E}_{p(\theta|z)}[\theta] = \operatorname{argmin}_{f(z)} \mathbb{E}_{p(z, \theta)} [\|f(z) - \theta\|^2] \quad (25)$$

614 *Proof.* Let  $\mathcal{L}[f]$  be the loss function defined as follows:

$$\mathcal{L}[f] = \mathbb{E}_{p(z)} [\|f(z) - \theta^*(z)\|^2]$$

615 Clearly, the minimizer of  $\mathcal{L}[f]$  is  $\theta^*(z)$ . With a bit of rearranging and using Bayes rule, we can  
 616 rewrite  $\mathcal{L}[f]$  as follows:

$$\begin{aligned} \mathcal{L}[f] &= \mathbb{E}_{p(z)} [\|f(z) - \theta^*(z)\|^2] \\ &= \mathbb{E}_{p(z)} [\|f(z)\|^2] - 2\mathbb{E}_{p(z)} [\langle f(z), \theta^*(z) \rangle] + \underbrace{\mathbb{E}_{p(z)} [\|\theta^*(z)\|^2]}_{\text{const. w.r.t. } f} \\ &= \mathbb{E}_{p(z, \theta)} [\|f(z)\|^2] - 2\mathbb{E}_{p(z)} [\langle f(z), \mathbb{E}_{p(\theta|z)}[\theta] \rangle] + \text{const.} \\ &= \mathbb{E}_{p(z, \theta)} [\|f(z)\|^2] - 2\mathbb{E}_{p(z, \theta)} [\langle f(z), \theta \rangle] + \text{const.} \\ &\quad \text{(complete the square)} \\ &= \mathbb{E}_{p(z, \theta)} [\|f(z) - \theta\|^2] - \underbrace{\mathbb{E}_{p(z, \theta)} [\|\theta\|^2]}_{\text{const. w.r.t. } f} + \text{const.} \end{aligned}$$

617 The minimizer of  $\mathcal{L}[f]$  is unaffected by the constant terms, and so we have that  $\theta^*(z) = \mathbb{E}_{p(\theta|z)}[\theta]$  is  
 618 the solution to

$$\operatorname{argmin}_{f(z)} \mathbb{E}_{p(z, \theta)} [\|\theta - f(z)\|^2]$$

619

□

## D Message passing

In this section we will review message passing and identify the key operations that are needed to perform message passing updates. We defer the discussion of numerically stable implementations of these operations to Appendix H. First we'll identify the key operations that are needed to perform message passing updates for the backward messages and then show how these operations can be used to perform message passing updates for the forward messages.

At a high level, the sequential and parallel message passing algorithms are variable elimination algorithms that eliminate different variables of the chain structured graph. The sequential algorithm operates on individual nodes and begins at one of the ends of the chain and sequentially eliminate variable at the end of the chain, whereas the parallel algorithm operates on pairs of nodes and eliminates the middle variable of the pair. For example, a rough sketch of the sequential elimination process looks like  $(0), 1, 2, 3, 4 \rightarrow (1), 2, 3, 4 \rightarrow (2), 3, 4 \rightarrow (3), 4 \rightarrow (4)$ , where the parentheses indicate the current node that is being processed. On the other hand, the parallel algorithm looks like  $(0, 1), 2, 3, 4 \rightarrow (0, 2), 3, 4 \rightarrow (0, 3), 4 \rightarrow (0, 4)$ .

### D.1 Sequential message passing

The sequential message passing updates for the backward messages can be written using the following recurrence relation:

$$\phi(x_{k-1}|\beta_{k-1}) = \int \phi_{k|k-1}(x_k|x_{k-1})\phi(x_k|\theta_k)\phi(x_k|\beta_k)dx_k, \quad \beta_N = 0 \quad (26)$$

See Appendix H.3 for pseudocode. There are two operations on Gaussians that are needed to perform these updates. The first is a “multiply” operation that takes two potential functions and returns a new potential function, and the second is an “update” operation that absorbs a potential function into a transition function.

**Definition 4 (Multiply).** Let  $\phi_1(x)$  and  $\phi_2(x)$  be potential functions over the same variable. Then the “multiply” operation is defined as

$$\phi_1(x)\phi_2(x) \mapsto \hat{\phi}(x) \quad (27)$$

When  $\phi_1(x)$  and  $\phi_2(x)$  are parameterized using natural parameters, then the multiply operation simply adds the natural parameters, i.e. if  $\theta_1$  and  $\theta_2$  are the natural parameters of  $\phi_1(x)$  and  $\phi_2(x)$ , then  $\phi_1(x|\theta_1)\phi_2(x|\theta_2) \mapsto \phi_1(x|\theta_1 + \theta_2)$ . We used this property to write the sequential message passing updates for the backward messages ?? We do note that when one uses a different parameterization, the multiply operation may look different. We will examples of this in Appendix H.

The second operation is the “update” operation, which absorbs a potential function into a transition function. This operation is what handles the integral in the recurrence relation.

**Definition 5 (Update).** Let  $\phi(y|x)$  be a transition function and  $\phi(y)$  be a potential function over the first variable. Then the “update” operation is defined as

$$\phi(y)\phi_{y|x}(y|x) \mapsto \hat{\phi}_{y|x}(y|x)\hat{\phi}(x) \quad (28)$$

where  $\hat{\phi}_{y|x}(y|x)$  and  $\hat{\phi}(x)$  are a new transition function and potential function, respectively.

Essentially, the update operation performs a change of variables of the coupling of  $x$  and  $y$  on the LHS. Furthermore, when the terms of the LHS are Gaussian, then the terms of the RHS are also Gaussian. This allows us to perform the update operation in closed form (see Appendix H).

The multiply and update operations are sufficient to perform the sequential message passing updates for the backward messages. For example, the backward message passing updates can be written as:

$$\int \phi_{k|k-1}(x_k|x_{k-1}) \underbrace{\phi(x_k|\theta_k)\phi(x_k|\beta_k)}_{\text{multiply} \rightarrow \phi(x_k|\theta_k+\beta_k)} dx_k \quad (29)$$

$$= \int \underbrace{\phi(x_k|\theta_k+\beta_k)\phi_{k|k-1}(x_k|x_{k-1})}_{\text{update} \rightarrow \hat{\phi}_{k|k-1}(x_k|x_{k-1})\phi(x_{k-1}|\beta_{k-1})} dx_k \quad (30)$$

$$= \underbrace{\int \hat{\phi}_{k|k-1}(x_k|x_{k-1})dx_k}_{\text{transition integrates to 1}} \phi(x_{k-1}|\beta_{k-1}) \quad (31)$$

$$= \phi(x_{k-1}|\beta_{k-1}) \quad (32)$$

658 The forward messages can be computed in a similar manner. The forward messages are given by:

$$\phi(x_{k+1}|\alpha_{k+1}) = \int \phi_{k+1|k}(x_{k+1}|x_k) \phi(x_k|\theta_k) \phi(x_k|\alpha_k) dx_k, \quad \alpha_1 = 0 \quad (33)$$

659 To find the forward messages, we can exploit the fact that our transition functions are Gaussian and  
 660 can therefore be reversed. This means that given a transition  $\phi(y|x)$ , we can find a reversed transition  
 661  $\phi^T(x|y)$  that evaluates to the same value as  $\phi(y|x)$  for all  $x, y$

662 **Definition 6** (Reversed transition). *Let  $\phi(y|x)$  be a transition function. Then the reversed transition*  
 663 *is defined as*

$$\phi^T(x|y) = \phi(y|x) \quad (34)$$

664 so that  $\phi^T(x|y) = \phi(y|x)$  for all  $x, y$  and  $\int \phi^T(x|y)dx = \int \phi(y|x)dx = 1$ .

665 Using this reverse operation, we can simply reverse the transition distributions and then find the  
 666 forward messages by using the same recurrence relation as for the backward messages:

$$\int \underbrace{\phi_{k+1|k}(x_{k+1}|x_k)}_{\text{reverse}} \underbrace{\phi(x_k|\theta_k)\phi(x_k|\alpha_k)}_{\text{multiply} \rightarrow \phi(x_k|\theta_k+\alpha_k)} dx_k \quad (35)$$

$$= \int \underbrace{\phi^T(x_k|x_{k+1})\phi(x_k|\theta_k+\alpha_k)}_{\text{update} \rightarrow \hat{\phi}^T(x_k|x_{k+1})\phi(x_{k+1}|\alpha_{k+1})} dx_k \quad (36)$$

$$= \underbrace{\int \hat{\phi}^T(x_k|x_{k+1})dx_k}_{\text{transition integrates to 1}} \phi(x_{k+1}|\alpha_{k+1}) \quad (37)$$

$$= \phi(x_{k+1}|\alpha_{k+1}) \quad (38)$$

667 These message passing updates can be computed in  $O(N)$  time using the the multiply, update and  
 668 reverse operations. However, there is a more efficient way to compute the forward messages using  
 669 the parallel scan algorithm [Särkkä and García-Fernández 2020] that reduces the complexity to  
 670  $O(\log N)$  on parallel compute. We will describe this algorithm in Appendix D.2

## 671 D.2 Parallel message passing

672 In this section we will use slightly different notation to describe the parallel message passing  
 673 algorithm. We will avoid writing out the parameters of our potential functions and call them by their  
 674 parameter name. For example, instead of writing  $\phi(x_k|\theta_k)$ , we will write  $\phi_k(x_k)$  and instead of  
 675 writing  $\phi(x_k|\beta_k)$ , we will write  $\beta(x_k)$ .

676 The building block of the parallel message passing algorithm [Särkkä and García-Fernández 2020] is  
 677 an unnormalized potential function over two variables, which we denote by  $\Psi(y, x)$ . We assume that  
 678  $\Psi(y, x)$  can be decomposed into a (normalized) transition distribution and an unnormalized potential  
 679 function:

$$\Psi(y, x) = \Psi(y|x)\Psi(x) \quad (39)$$

680 Whenever we write  $\Psi(y|x)$ , we are referring to a valid conditional probability distribution  
 681 ( $\int \Psi(y|x)dy = 1$ ). Since  $\Psi(y, x)$  is jointly Gaussian over  $x$  and  $y$ , we are able to integrate out  
 682 variables in  $x$  and  $y$  and can also combine neighboring potentials into a new Gaussian potential.  
 683 These properties allow us to construct a chain operation over potentials that combines neighboring  
 684 potentials and then integrates out the common variable. We denote this chain operation by  $\otimes$ :

$$\Psi(y, x) := \int \Psi(y, z)\Psi(z, x)dz =: \Psi(y, z) \otimes \Psi(z, x) \quad (40)$$

685 An important property of the chain operation is that it is associative due to the fact that we can swap  
 686 the order or integration (we will prove this in Appendix D.3).

687 A useful perspective of this chain operation is that it amounts to performing variable elimination on  
 688 the graph defined by the potentials, i.e. performs some sort of message passing [Koller, 2009]. With  
 689 this in mind, we can perform message passing by constructing the appropriate joint potentials:

690 **Proposition 9** (Parallel messages). *Let  $\phi_{k+1|k}$  and  $\phi_k$  be the potential functions for the CRF in*  
 691 *Definition 1 and  $\alpha$  and  $\beta$  be the messages defined in Eqs. (26) and (33). Then*

$$\alpha_k(x_k) = \int \Psi_{1:k}^{fwd}(x_k, x_1) dx_1 \quad \text{and} \quad \beta_k(x_k) = \int \Psi_{k:N}^{bwd}(x_N | x_k) dx_N \quad (41)$$

692 where

$$\Psi_{1:k}^{fwd}(x_k, x_1) = \bigotimes_{i=1}^{k-1} \phi_{i+1|i}(x_{i+1} | x_i) \phi_i(x_i) \quad (42)$$

$$\text{and} \quad \Psi_{k:N}^{bwd}(x_N | x_k) = \bigotimes_{i=N-1}^k \phi_{i+1|i}(x_{i+1} | x_i) \phi_{i+1}(x_{i+1}) \quad (43)$$

693 See appendix Appendix D.3 for a proof and ?? for pseudocode. Since  $\otimes$  is associative, we can  
 694 evaluate Eq. (42) in  $O(\log N)$  time using the parallel scan algorithm [Särkkä and García-Fernández  
 695 2020]. The rough idea is that on parallel compute, one can, in parallel, chain together consecutive  
 696 pairs of potentials and then recurse on these new chained potentials in order to eventually chain the  
 697 entire sequence. We provide pseudocode for this a special case of this algorithm in Appendix H.3  
 698  $\Psi_{1:k}^{fwd}(x_k, x_1)$  and  $\Psi_{k:N}^{bwd}(x_N | x_k)$  can be thought of as the result of marginalization over the variables  
 699 between  $x_1$  and  $x_k$  and  $x_k$  and  $x_N$ , respectively.

### 700 D.3 Chain operation

701 Recall that the chain operation is defined in Eq. (40) as

$$\Psi(y, x) := \int \Psi(y, z) \Psi(z, x) dz =: \Psi(y, z) \otimes \Psi(z, x) \quad (44)$$

702 To see that it is associative, we need to check that  $\Psi(y, z) \otimes (\Psi(z, x) \otimes \Psi(x, w)) =$   
 703  $(\Psi(y, z) \otimes \Psi(z, x)) \otimes \Psi(x, w)$

$$\Psi(y, z) \otimes (\Psi(z, x) \otimes \Psi(x, w)) = \int \Psi(y, z) \left( \int \Psi(z, x) \Psi(x, w) dx \right) dz \quad (45)$$

$$= \int \int \Psi(y, z) \Psi(z, x) \Psi(x, w) dx dz \quad (46)$$

$$= \int \left( \int \Psi(y, z) \Psi(z, x) dz \right) \Psi(x, w) dx \quad (47)$$

$$= (\Psi(y, z) \otimes \Psi(z, x)) \otimes \Psi(x, w) \quad (48)$$

704 **Proposition 10** (Parallel messages). *Let  $\phi_{k+1|k}$  and  $\phi_k$  be the potential functions for the CRF in*  
 705 *Definition 1 and  $\alpha$  and  $\beta$  be the messages defined in Eqs. (26) and (33). Then*

$$\alpha_k(x_k) = \int \Psi_{1:k}^{fwd}(x_k, x_1) dx_1 \quad \text{and} \quad \beta_k(x_k) = \int \Psi_{k:N}^{bwd}(x_N | x_k) dx_N \quad (49)$$

706 where

$$\Psi_{1:k}^{fwd}(x_k, x_1) = \bigotimes_{i=1}^{k-1} \phi_{i+1|i}(x_{i+1} | x_i) \phi_i(x_i) \quad (50)$$

$$\text{and} \quad \Psi_{k:N}^{bwd}(x_N | x_k) = \bigotimes_{i=N-1}^k \phi_{i+1|i}(x_{i+1} | x_i) \phi_{i+1}(x_{i+1}) \quad (51)$$

707 *Proof.* First for notational clarity, define

$$\Psi_{i+1,i}^{\text{bwd}}(x_{i+1}|x_i) = \phi_{i+1|i}(x_{i+1}|x_i)\phi_{i+1}(x_{i+1}) \quad \text{and} \quad \Psi_{i+1,i}^{\text{fwd}}(x_{i+1}, x_i) = \phi_{i+1|i}(x_{i+1}|x_i)\phi_i(x_i) \quad (52)$$

708 We can compute the cumulative potentials as follows:

$$\Psi_{k:N}^{\text{bwd}}(x_N|x_k) = \bigotimes_{i=N-1}^k \Psi_{i+1,i}^{\text{bwd}}(x_{i+1}|x_i) \quad (53)$$

$$= \Psi_{N:N-1}^{\text{bwd}}(x_N|x_{N-1}) \otimes \Psi_{N-1:N-2}^{\text{bwd}}(x_{N-1}|x_{N-2}) \otimes \cdots \otimes \Psi_{k+1:k}^{\text{bwd}}(x_{k+1}|x_k) \quad (54)$$

$$= \int \Psi_{N:N-1}^{\text{bwd}}(x_N|x_{N-1}) \int \Psi_{N-1:N-2}^{\text{bwd}}(x_{N-1}|x_{N-2}) dx_{N-1} \int \Psi_{N-2:N-3}^{\text{bwd}}(x_{N-2}|x_{N-3}) dx_{N-2} \cdots dx_{k+1} \quad (55)$$

$$= \int \cdots \int \prod_{i=k}^{N-1} \Psi_{i+1,i}^{\text{bwd}}(x_{i+1}|x_i) dx_{N-1} \cdots dx_{k+1} \quad (56)$$

709 And similarly for the forward potentials:

$$\Psi_{1:k}^{\text{fwd}}(x_k, x_1) = \bigotimes_{i=1}^{k-1} \Psi_{i+1,i}^{\text{fwd}}(x_{i+1}, x_i) \quad (57)$$

$$= \int \cdots \int \prod_{i=1}^{k-1} \Psi_{i+1,i}^{\text{fwd}}(x_{i+1}, x_i) dx_2 \cdots dx_{k-1} \quad (58)$$

710 Next, we can rewrite the joint distribution of the CRF in a similar form:

$$p(x_{1:N}) = \prod_{k=1}^{N-1} \phi_{k+1|k}(x_{k+1}|x_k) \prod_{k=1}^N \phi_k(x_k) \quad (59)$$

$$= \phi_k(x_k) \prod_{i=k}^{N-1} \Psi_{i+1,i}^{\text{bwd}}(x_{i+1}|x_i) \prod_{i=1}^{k-1} \Psi_{i+1,i}^{\text{fwd}}(x_{i+1}, x_i), \quad \forall k \in \{1, \dots, N\} \quad (60)$$

711 Then, integrating over the variables  $dx_1, \dots, \hat{dx}_k, \dots, dx_N$ , where  $\hat{dx}_k$  denotes that we are not  
712 integrating over  $x_k$ , completes the proof:

$$p(x_k) = \int \cdots \int p(x_{1:N}) dx_1 \dots \hat{dx}_k \dots dx_N \quad (61)$$

$$\propto \int \cdots \int \prod_{k=1}^{N-1} \phi_{k+1|k}(x_{k+1}|x_k) \prod_{k=1}^N \phi_k(x_k) dx_1 \dots \hat{dx}_k \dots dx_N \quad (62)$$

$$= \phi_k(x_k) \int \cdots \int \prod_{i=k}^{N-1} \Psi_{i+1,i}^{\text{bwd}}(x_{i+1}|x_i) \prod_{i=1}^k \Psi_{i+1,i}^{\text{fwd}}(x_{i+1}, x_i) dx_1 \dots \hat{dx}_k \dots dx_N \quad (63)$$

$$= \phi_k(x_k) \underbrace{\int \Psi_{k:N}^{\text{bwd}}(x_N|x_k) dx_N}_{\beta_k(x_k)} \underbrace{\int \Psi_{1:k}^{\text{fwd}}(x_k, x_1) dx_1}_{\alpha_k(x_k)} \quad (64)$$

713 We can recognize the terms in the last equation as the forward and backward messages, which  
714 completes the proof.  $\square$

715 It will be convenient later to define an operator that actually transforms the parameters of the backward  
716 messages.

717 **Definition 7** (Message passing update operator). Let  $\phi_{k+1|k}(x_{k+1}, x_k)$  be a Gaussian transition  
 718 function and let  $\phi(x_{k+1}|\eta_{k+1})$  be a Gaussian node potential with natural parameters  $\eta_{k+1}$ . Next  
 719 consider the message passing update:

$$\phi(x_k|\eta_k) = \int \phi_{k+1|k}(x_{k+1}|x_k)\phi(x_{k+1}|\eta_{k+1})dx_{k+1} \quad (65)$$

720 The message passing update operator is denoted by  $\Phi_{k,k+1}(\eta_{k+1})$  and is defined to satisfy:

$$\eta_k = \Phi_{k,k+1}(\eta_{k+1}) \quad (66)$$

721 In particular, the update rule for the backward messages is given by:

$$\beta_k = \Phi_{k,k+1}(\beta_{k+1} + \theta_{k+1}) \quad (67)$$

722 **Corollary 2** (Mixed parameterization update rule). Let  $\phi_{k+1|k}(x_{k+1}|x_k) := N(x_{k+1}|Ax_k + u, \Sigma)$  be  
 723 a Gaussian transition function and let  $\phi(x_{k+1}|\eta_{k+1}) := N(x_{k+1}|\mu_{k+1}, J_{k+1}^{-1})$  be a Gaussian node  
 724 potential where  $J_{k+1}$  is the precision matrix. If  $\eta_k$  and  $\eta_{k+1}$  represent the mean and precision matrix  
 725 of a Gaussian distribution, then the update and marginalize operator is denoted by  $\Phi_{k,k+1}(\eta_{k+1})$   
 726 and is given by:

$$\Phi_{k,k+1}(\mu_{k+1}, J_{k+1}) = (A^{-1}(\mu_{k+1} - u), \Phi_{k,k+1}^{(J)}(J_{k+1})) \quad (68)$$

727 where  $\Phi_{k,k+1}^{(J)}(J_{k+1})$  is a nonlinear function of  $J_{k+1}$ .

728 *Proof.* The result follows from Appendix [H.3](#). □

#### 729 D.4 Probabilistic queries

730 The forward and backward messages can be used to compute the majority of the probabilistic queries  
 731 of interest on a CRF. Recall our definition of a CRF:

$$p(x_{1:N}|\theta) \propto \prod_{k=1}^{N-1} \phi_{k+1|k}(x_{k+1}|x_k) \prod_{k=1}^N \phi(x_k|\theta_k) \quad (69)$$

732 Next we will describe two probabilistic queries of interest: the marginal distribution and the transition  
 733 distribution.

**Proposition 11** (Marginal distribution).

$$p(x_k|\theta) = \phi(x_k|\theta_k + \alpha_k + \beta_k) \quad (70)$$

734 *Proof.* The derivation is given in Eq. [\(61\)](#). For completeness, we will change notation:

$$p(x_k) = \phi_k(x_k)\beta_k(x_k)\alpha_k(x_k) \text{ (notation in previous section)} \quad (71)$$

$$:= \phi(x_k|\theta_k)\phi(x_k|\alpha_k)\phi(x_k|\beta_k) \text{ (notation in this section and in main text)} \quad (72)$$

$$= \phi(x_k|\theta_k + \alpha_k + \beta_k) \quad (73)$$

735 □

**Proposition 12** (Transition distribution).

$$p(x_{k+1}|x_k, \theta) \propto \phi_{k+1|k}(x_{k+1}|x_k)\phi(x_{k+1}|\theta_{k+1} + \beta_{k+1}) \quad (74)$$

736 *Proof.* We can start by computing the joint distribution  $p(x_{k+1}, x_k|\theta)$ . By using variable elimination,  
 737 we can show that

$$p(x_{k+1}, x_k|\theta) = \phi(x_k|\alpha_k)\phi_{k+1|k}(x_{k+1}|x_k)\phi(x_{k+1}|\theta_{k+1})\phi(x_{k+1}|\beta_{k+1}) \quad (75)$$

738 Dividing by the marginal distribution  $p(x_k|\theta)$  and using the definition of the transition distribution,  
 739 we get

$$p(x_{k+1}|x_k, \theta) = \phi_{k+1|k}(x_{k+1}|x_k) \frac{\phi(x_{k+1}|\beta_{k+1} + \theta_{k+1})}{\phi(x_k|\beta_k + \theta_k)} \quad (76)$$

740 which, after absorbing the denominator into the normalization constant, is equivalent to the desired  
 741 result. □



742 **Corollary 3** (Autoregressive factorization). *The autoregressive factorization of  $p(x_{1:N}|\theta)$  takes the*  
 743 *following form:*

$$p(x_{1:N}|\theta) \propto \phi(x_1|\theta_1 + \beta_1) \prod_{k=1}^{N-1} \phi_{k+1|k}(x_{k+1}|x_k) \phi(x_{k+1}|\theta_{k+1} + \beta_{k+1}) \quad (77)$$

744 *Proof.* This follows directly from applying Proposition 11 and Proposition 12 to  $p(x_{1:N}|\theta) =$   
 745  $p(x_1|\theta) \prod_{k=1}^{N-1} p(x_{k+1}|x_k, \theta)$ .  $\square$

## 746 E Conditioned SDEs

747 In this section we derive the form of conditioned linear SDEs as well as the corresponding probability  
 748 flow ODEs.

### 749 E.1 Conditioned linear SDE

750 **Proposition 13** (Conditioned Linear SDE). *Let  $\phi_{t+s|t}(x_{t+s}|x_t)$  be the transition distribution of the*  
 751 *linear SDE  $dx_t = F_t x_t dt + L_t dW_t$  and let  $\{\phi(x_{t_k}|\theta_{t_k})\}_{t_k \in \mathcal{R}}$  be potential functions at times in the*  
 752 *set  $\mathcal{R}$ . Then the piecewise-linear SDE,*

$$dx_t = (F_t x_t + L_t L_t^T \nabla \log \phi(x_t|\beta_t)) dt + L_t dW_t, \quad x_{t_1} \sim \phi(x_{t_1}|\beta_1 + \theta_1) \quad (78)$$

753 *where  $t \in (t_k, t_{k+1})$  and  $t_k, t_{k+1} \in \mathcal{R}$ , has a joint distribution over any superset of times  $t_{1:N} =$*   
 754  *$\mathcal{T} \supseteq \mathcal{R}$  that is given by a CRF:*

$$p(x_{t_{1:N}}|\theta) \propto \prod_{t_k \in \mathcal{T}} \phi_{t_{k+1}|t_k}(x_{t_{k+1}}|x_{t_k}) \prod_{t_k \in \mathcal{R}} \phi(x_{t_k}|\theta_{t_k}) \quad (79)$$

755 *where  $\beta_t$  is the extension of the backward message defined in ?? to time  $t$ :*

$$\phi(x_t|\beta_t) = \int \phi_{t_{k+1}|t}(x_{t_{k+1}}|x_t) \phi(x_{t_{k+1}}|\theta_{t_{k+1}} + \beta_{t_{k+1}}) dx_{t_{k+1}} \quad (80)$$

756 *Proof.* We will first construct the transition distribution of the conditioned SDE and then use Doob's  
 757 *h-transform to identify the form of the SDE. Recall that Doob's h-transform ([Särkkä and Solin*  
 758 *2019] section 7.5) is used to find the SDE associated with a transition distribution of the form*  
 759  *$p(x_{t+s}|x_t) = \phi_{t+s|t}(x_{t+s}|x_t) \frac{h_{t+s}(x_{t+s})}{h_t(x_t)}$  where  $\phi_{t+s|t}(x_{t+s}|x_t)$  is the transition distribution of*  
 760 *a base SDE with the form  $dx_t = u_t dt + L_t dW_t$  and  $h_t$  is a function that satisfies  $h_t(x_t) =$*   
 761  *$\int_t^{t+s} \phi_{t+s|t}(x_{t+s}|x_t) h_{t+s}(x_{t+s}) dx_{t+s}$ . Then the SDE whose transition distribution is  $p(x_{t+s}|x_t)$  is*  
 762 *given by*

$$dx_t = (u_t + L_t L_t^T \nabla \log h_t(x_t)) dt + L_t dW_t \quad (81)$$

763 We will show that the backward messages of the CRF are of the form  $h_t(x_t)$  and then use Doob's  
 764 h-transform to identify the form of the conditioned SDE.

765 Suppose  $t \in (t_k, t_{k+1})$  and  $s > 0$  is small enough so that  $t + s \in (t_k, t_{k+1})$ . Then we can construct  
 766 the joint distribution over  $(t_{t+s}, t_{k+1}, \dots, t_N)$  given  $x_t$  as

$$p(x_{t+s}|x_t) = \int \cdots \int p(x_{t_{k+1:N}}, x_{t+s}|x_t) dx_{t_{k+1}} \cdots dx_{t_N} \quad (82)$$

$$\propto \int \cdots \int \phi(x_{t_{k+1}}|\theta_{t_{k+1}}) \underbrace{\left( \prod_{i=k+1}^{N-1} \phi_{t_{i+1}|t_i}(x_{t_{i+1}}|x_{t_i}) \phi(x_{t_{i+1}}|\theta_{t_{i+1}}) \right)}_{\text{integrate to get parallel bwd message (Proposition 9)}} \phi_{t_{k+1}|t+s}(x_{t_{k+1}}|x_{t+s}) dx_{t_{k+1}} \cdots dx_{t_N} \phi_{t+s|t}(x_{t+s}|x_t) \quad (83)$$

$$= \int \int \phi(x_{t_{k+1}}|\theta_{t_{k+1}}) \Psi_{k+1:N}^{\text{bwd}}(x_{t_N}|x_{t_{k+1}}) \phi_{t_{k+1}|t+s}(x_{t_{k+1}}|x_{t+s}) dx_{t_N} dx_{t_{k+1}} \phi_{t+s|t}(x_{t+s}|x_t) \quad (84)$$

$$= \underbrace{\int \phi(x_{t_{k+1}}|\theta_{t_{k+1}}) \phi(x_{t_{k+1}}|\beta_{t_{k+1}}) \phi_{t_{k+1}|t+s}(x_{t_{k+1}}|x_{t+s}) dx_{t_{k+1}}}_{=: \phi(x_{t+s}|\beta_{t+s})} \phi_{t+s|t}(x_{t+s}|x_t) \quad (85)$$

$$= \phi(x_{t+s}|\beta_{t+s})\phi_{t+s|t}(x_{t+s}|x_t) \quad (86)$$

767 We can find the normalizing constant by integrating over  $x_{t+s}$ :

$$\int \phi(x_{t+s}|\beta_{t+s})\phi_{t+s|t}(x_{t+s}|x_t)dx_{t+s} \quad (87)$$

$$= \int \int \phi(x_{t_{k+1}}|\theta_{t_{k+1}})\phi(x_{t_{k+1}}|\beta_{t_{k+1}})\phi_{t_{k+1}|t+s}(x_{t_{k+1}}|x_{t+s})dx_{t_{k+1}}\phi_{t+s|t}(x_{t+s}|x_t)dx_{t+s} \quad (88)$$

$$= \int \phi(x_{t_{k+1}}|\theta_{t_{k+1}})\phi(x_{t_{k+1}}|\beta_{t_{k+1}}) \underbrace{\int \phi_{t_{k+1}|t+s}(x_{t_{k+1}}|x_{t+s})\phi_{t+s|t}(x_{t+s}|x_t)dx_{t+s}}_{\phi_{t_{k+1}|t}(x_{t_{k+1}}|x_t)} dx_{t_{k+1}} \quad (89)$$

$$= \int \phi(x_{t_{k+1}}|\theta_{t_{k+1}})\phi(x_{t_{k+1}}|\beta_{t_{k+1}})\phi_{t_{k+1}|t}(x_{t_{k+1}}|x_t)dx_{t_{k+1}} \quad (90)$$

$$= \phi(x_t|\beta_t) \quad (91)$$

768 Therefore, the transition distribution is

$$p(x_{t+s}|x_t) = \phi_{t+s|t}(x_{t+s}|x_t) \frac{\phi(x_{t+s}|\beta_{t+s})}{\phi(x_t|\beta_t)} \quad (92)$$

769 Note that Eq. (87) also verifies that  $\phi(x_t|\beta_t)$  satisfies the normalization condition for  $h_t(x_t)$  in Doob's  
770 h-transform. Directly applying Doob's h-transform to the transition distribution in Eq. (82) identifies  
771 the form of the conditioned SDE:

$$dx_t = (F_t x_t + L_t L_t^T \nabla \log \phi(x_t|\beta_t))dt + L_t dW_t \quad (93)$$

772 This piecewise-linear SDE has the correct conditional distribution,  $p(x_t|x_{t_{k_1}})$ , but requires an initial  
773 distribution. One can verify that the initial distribution  $p(x_{t_1}) \propto \phi(x_{t_1}|\theta_{t_1} + \beta_{t_1})$  is the first marginal  
774 distribution of the CRF in Definition 1  $\square$

## 775 E.2 Probabilistic queries for conditioned linear SDEs

776 **Lemma 2** (Marginal distribution of conditioned SDE). *Suppose  $t \in (t_k, t_{k+1})$  is a time in between*  
777 *the inducing points  $t_k$  and  $t_{k+1}$  of the conditioned linear SDE in Proposition 4. Then the marginal*  
778 *distribution of the SDE at time  $t$  is given by*

$$p(x_t) = \phi(x_t|\alpha_t + \beta_t) \quad (94)$$

779 where  $\alpha_t$  and  $\beta_t$  are extensions of the forward and backward messages defined in Eq. (33) and  
780 Eq. (26) to time  $t$ :

$$\phi(x_t|\alpha_t) = \int \phi_{t|t_{k-1}}(x_t|x_{t_{k-1}})\phi(x_{t_{k-1}}|\theta_{t_{k-1}} + \alpha_{t_{k-1}})dx_{t_{k-1}} \quad (95)$$

781 and

$$\phi(x_t|\beta_t) = \int \phi_{t|t_{k+1}}(x_t|x_{t_{k+1}})\phi(x_{t_{k+1}}|\theta_{t_{k+1}} + \beta_{t_{k+1}})dx_{t_{k+1}} \quad (96)$$

782 *Proof.* We can simply incorporate  $t$  into the set discretization times,  $t_{1:N}$ , used in Proposition 4 to  
783 get the desired result. Suppose  $t \in (t_i, t_{i+1})$  for some  $i$ . Then we can write the joint distribution as

$$p(x_t, x_{t_{1:N}}|\theta) \propto \phi_{t_{i+1}|t_i}(x_{t_{i+1}}|x_{t_i})\phi_{t|t_i}(x_t|x_{t_i}) \prod_{t_k \in \mathcal{T}} \phi_{t_{k+1}|t_k}(x_{t_{k+1}}|x_{t_k}) \prod_{t_k \in \mathcal{R}} \phi(x_{t_k}|\theta_{t_k}) \quad (97)$$

784 Then we can run variable elimination on the ends of the chain until we are left with the marginal  
785 distribution of  $x_t$ :

$$p(x_t) = \int p(x_t, x_{t_{1:N}}|\theta)dx_{t_{1:N}} \quad (98)$$

$$= \int \int \phi(x_{t_i}|\alpha_{t_i} + \theta_{t_i})\phi_{t|t_i}(x_t|x_{t_i})\phi_{t_{i+1}|t}(x_{t_{i+1}}|x_t)\phi(x_{t_{i+1}}|\beta_{t_{i+1}} + \theta_{t_{i+1}})dx_{t_{i+1}}dx_{t_i} \quad (99)$$

$$= \underbrace{\int \phi(x_{t_i}|\alpha_{t_i} + \theta_{t_i}) \phi_{t|t_i}(x_t|x_{t_i}) dx_{t_i}}_{\phi(x_t|\alpha_t)} \underbrace{\int \phi_{t_{i+1}|t}(x_{t_{i+1}}|x_t) \phi(x_{t_{i+1}}|\beta_{t_{i+1}} + \theta_{t_{i+1}}) dx_{t_{i+1}}}_{\phi(x_t|\beta_t)} \quad (100)$$

$$= \phi(x_t|\alpha_t + \beta_t) \quad (101)$$

786

787 **Lemma 3** (Transition distribution of conditioned linear SDE). *Suppose  $t \in (t_k, t_{k+1})$  is a time in*  
 788 *between the inducing points  $t_k$  and  $t_{k+1}$  of the conditioned linear SDE in Proposition 4 and suppose*  
 789 *that  $s > 0$  is small enough so that  $t + s \in (t_k, t_{k+1})$ . Then the transition distribution of the SDE at*  
 790 *time  $t$  is given by*

$$\phi_{t+s|t}(x_{t+s}|x_t) \propto \phi_{t+s|t}(x_{t+s}|x_t) \phi(x_{t+s}|\beta_{t+s}) \quad (102)$$

791 *Proof.* The proof is embedded in the derivation of the conditioned linear SDE at Eq. (92).  $\square$

792 **Corollary 4** (Autoregressive factorization). *The autoregressive factorization of  $p(x_{t_{1:N}}|\theta)$  is given*  
 793 *by*

$$p(x_{t_{1:N}}|\theta) = p(x_{t_1}|\theta) \prod_{t_k \in \mathcal{T}} \phi_{t_k|t_{k-1}}(x_{t_k}|x_{t_{k-1}}) \phi(x_{t_k}|\beta_{t_k}) \quad (103)$$

$$\text{where } \beta_{t_k} = \begin{cases} \Phi_{t_k, t_{k+1}}(\beta_{t_{k+1}} + \theta_{t_{k+1}}) & \text{if } t_k \in \mathcal{R} \\ \Phi_{t_k, t_{k+1}}(\beta_{t_{k+1}}) & \text{otherwise} \end{cases} \quad (104)$$

794 *where  $\Phi_{t_k, t_{k+1}}$  is the message passing update operator defined in Definition 7*

795 *Proof.* Recall that

$$p(x_{t_{1:N}}|\theta) \propto \prod_{t_k \in \mathcal{T}} \phi_{t_{k+1}|t_k}(x_{t_{k+1}}|x_{t_k}) \prod_{t_k \in \mathcal{R}} \phi(x_{t_k}|\theta_{t_k}) \quad (105)$$

796 Suppose that for each  $t_k \notin \mathcal{R}$ , we introduce a new potential function whose natural parameters are 0,  
 797 which we will denote by  $\phi(x_{t_k}|\emptyset_{t_k})$ . These new potentials have no effect on the joint distribution,  
 798 but allow us to rewrite the joint distribution in the same form as in Corollary 3 which yields the  
 799 result.  $\square$

### 800 E.3 Probability flow ODE for conditioned linear SDEs

801 **Corollary 5** (Probability flow ODE). *The probability flow ODE of the SDE in Proposition 4 is given*  
 802 *by*

$$\frac{dx_t}{dt} = F_t x_t + \frac{1}{2} L_t L_t^T (\nabla \log \phi(x_t|\beta_t) - \nabla \log \phi(x_t|\alpha_t)) \quad (106)$$

803  $\beta_t$  is the same as in Proposition 4 and  $\alpha_t$  is the extension of the forward message defined in Eq. (33)  
 804 to time  $t$ :

$$\phi(x_t|\alpha_t) = \int \phi_{t|t_k}(x_t|x_{t_k}) \phi(x_{t_k}|\theta_{t_k} + \alpha_{t_k}) dx_{t_k} \quad (107)$$

805 *Proof.* Let  $dx_t = u_t dt + L_t dW_t$  be an SDE. Then the probability flow ODE is defined Song et al.  
 806 [2021] as

$$\frac{dx_t}{dt} = u_t - \frac{1}{2} L_t L_t^T \nabla \log p_t(x_t) \quad (108)$$

807 where  $p_t(x_t)$  is defined as the marginal distribution of the SDE, which is given by Lemma 2. We can  
 808 apply this directly to our SDE in Proposition 4 to get the result:

$$\frac{dx_t}{dt} = (F_t x_t + L_t L_t^T \nabla \log \phi(x_t|\beta_t)) - \frac{1}{2} L_t L_t^T \nabla \log p_t(x_t) \quad (109)$$

$$= (F_t x_t + L_t L_t^T \nabla \log \phi(x_t|\beta_t)) - \frac{1}{2} L_t L_t^T (\nabla \log \phi(x_t|\alpha_t) + \nabla \log \phi(x_t|\beta_t)) \quad (110)$$

$$= F_t x_t + \frac{1}{2} L_t L_t^T (\nabla \log \phi(x_t|\beta_t) - \nabla \log \phi(x_t|\alpha_t)) \quad (111)$$

809  $\square$

## 810 F CMFVI proofs

### 811 F.1 Constrained mean field VI

812 Let  $\theta \sim p(\theta)$  be an unknown prior distribution on the parameters of the conditional exponential  
 813 family distribution,  $p(x|z, \theta) \propto \exp\{\langle t_z(x), \theta \rangle - A(z, \theta)\}$ , where  $t_z(x)$  is the sufficient statistic  
 814 of the exponential family distribution and  $A(z, \theta)$  is the log partition function. In our setting, we  
 815 interpret  $x$  and  $z$  as unobserved and observed variables and  $\theta$  as a parameter that they both depend  
 816 on. We are interested in performing inference in the predictive distribution  $p(x|z)$ , where we must  
 817 integrate out  $\theta$ . This distribution can be written as:

$$p(x|z) = \int p(x|z, \theta) p(\theta|z) d\theta \quad (112)$$

$$= \mathbb{E}_{p(\theta|z)} [\exp\{\langle t_z(x), \theta \rangle - A(z, \theta)\}] \quad (113)$$

818 where  $t_z(x)$  is the sufficient statistic of the conditional exponential family distribution. Since this  
 819 distribution is intractable, we use a variational approximation to approximate it. Our variational  
 820 approximation is called the constrained mean field VI approximation and is given by:

$$q^*(x|z) = \underset{q(x|z)}{\operatorname{argmin}} \operatorname{KL} [q(x|z)p(\theta|z) \| p(x, \theta|z)] \quad (114)$$

821 In this appendix section we will derive facts about  $q^*(x|z)$ .

822 **Lemma 4** (Alternate constrained mean field VI objectives). *The constrained mean field VI objective,*

$$\operatorname{KL} [q(x|z)p(\theta|z) \| p(x, \theta|z)] \quad (115)$$

823 *is equal to the following expressions:*

1.

$$\mathbb{E}_{q(x|z)p(\theta|z)} \left[ \log \frac{p(\theta|z)}{p(\theta|x, z)} \right] + \operatorname{KL} [q(x|z) \| p(x|z)] \quad (116)$$

2.

$$\mathbb{E}_{q(x|z)p(\theta|z)} \left[ \log \frac{p(x|z)}{p(x|z, \theta)} \right] + \operatorname{KL} [q(x|z) \| p(x|z)] \quad (117)$$

3.

$$\mathbb{E}_{q(x|z)} [\log q(x|z) - \mathbb{E}_{p(\theta|z)} [\log p(x|z, \theta)]] \quad (118)$$

824 *Proof.* The proof is a straightforward rearrangement of terms:

$$\operatorname{KL} [q(x|z)p(\theta|z) \| p(x, \theta|z)] = \int \int q(x|z)p(\theta|z) \log \frac{q(x|z)p(\theta|z)}{p(x, \theta|z)} dx dy \quad (119)$$

$$= \int \int q(x|z)p(\theta|z) \log \frac{p(\theta|z)}{p(\theta|x, z)} \frac{q(x|z)}{p(x|z)} dx dy \quad (\text{equals 1}) \quad (120)$$

$$= \int \int q(x|z)p(\theta|z) \log \frac{p(\theta|z)}{p(x|z, \theta)} \frac{q(x|z)}{p(x|z, \theta)} dx dy \quad (\text{equals 2}) \quad (121)$$

$$= \int \int q(x|z)p(\theta|z) \log \frac{q(x|z)}{p(x|z, \theta)} dx dy \quad (122)$$

$$= \mathbb{E}_{q(x|z)} [\log q(x|z) - \mathbb{E}_{p(\theta|z)} [\log p(x|z, \theta)]] \quad (123)$$

825  $\square$

826 **Theorem 2** (Constrained mean field VI solution). *Let  $p(x|z, \theta) \propto \exp\{\langle t_z(x), \theta \rangle - A(z, \theta)\}$  be an*  
 827 *exponential family distribution and that  $\theta \sim p(\theta|z)$ . The constrained mean field VI approximation of*  
 828  *$p(x|z)$ , denoted by  $q^*(x|z)$ , is defined as follows:*

$$q^*(x|z) = \underset{q(x|z)}{\operatorname{argmin}} \operatorname{KL} [q(x|z)p(\theta|z) \| p(x, \theta|z)] \quad (124)$$

$$= p(x|z, \theta^*(z)), \quad \text{where } \theta^*(z) = \mathbb{E}_{p(\theta|z)} [\theta] \quad (125)$$

829 *Proof.* The proof can follow quickly from the standard mean field VI solutions [Beal \[2003\]](#), but for  
 830 completeness we will derive it from scratch. Starting from the result of Lemma [4](#), we have that

$$q^*(x|z) = \operatorname{argmin}_{q(x|z)} \mathbb{E}_{q(x|z)} [\log q(x|z) - \mathbb{E}_{p(\theta|z)} [\log p(x|z, \theta)]] \quad (126)$$

831 We can introduce a Lagrange multiplier to enforce the constraint that the distribution is normalized.  
 832 Let  $q_\epsilon(x|z) = q(x|z) + \epsilon \eta(x|z)$  where  $\eta$  is the variation function and  $\epsilon$  is a scalar. Then we can take  
 833 a variation by differentiating with respect to  $\epsilon$ :

$$\frac{\partial}{\partial \epsilon} \left( \mathbb{E}_{q_\epsilon(x|z)} [\log q_\epsilon(x|z) - \mathbb{E}_{p(\theta|z)} [\log p(x|z, \theta)]] + \lambda \left( \int q_\epsilon(x|z) dx - 1 \right) \right) = 0 \quad (127)$$

$$\implies \frac{\partial}{\partial \epsilon} \int q_\epsilon(x|z) \log q_\epsilon(x|z) dx + \int \eta(x|z) (\mathbb{E}_{p(\theta|z)} [\log p(x|z, \theta)] + \lambda) dx = 0 \quad (128)$$

834 The negative entropy term simplifies as follows:

$$\frac{\partial}{\partial \epsilon} \int q_\epsilon(x|z) \log q_\epsilon(x|z) dx = \int \frac{\partial}{\partial \epsilon} q_\epsilon(x|z) \log q_\epsilon(x|z) dx + \int q_\epsilon(x|z) \frac{\partial}{\partial \epsilon} \log q_\epsilon(x|z) dx \quad (129)$$

$$= \int \frac{\partial q_\epsilon(x|z)}{\partial \epsilon} \log q_\epsilon(x|z) dx + \int q_\epsilon(x|z) \frac{\partial \log q_\epsilon(x|z)}{\partial \epsilon} dx \quad (130)$$

$$= \int \eta(x|z) \log q_\epsilon(x|z) dx - \int q_\epsilon(x|z) \frac{1}{q_\epsilon(x|z)} \frac{\partial q_\epsilon(x|z)}{\partial \epsilon} dx \quad (131)$$

$$= \int \eta(x|z) (\log q_\epsilon(x|z) - 1) dx \quad (132)$$

835 Plugging this back into the original equation and setting it equal to zero implies that the integrand  
 836 must be zero:

$$\mathbb{E}_{p(\theta|z)} [\log p(x|z, \theta)] + \lambda + \log q_\epsilon(x|z) - 1 = 0 \quad (133)$$

837 Solving for  $\log q_\epsilon(x|z)$  (and setting  $\epsilon = 0$ ) yields:

$$\log q(x|z) = \mathbb{E}_{p(\theta|z)} [\log p(x|z, \theta)] + \lambda - 1 \quad (134)$$

838 The lagrange multiplier  $\lambda$  ensures that the distribution is normalized, and so we have that

$$q^*(x|z) = \exp \{ \mathbb{E}_{p(\theta|z)} [\log p(x|z, \theta)] + \lambda - 1 \} \quad (135)$$

$$\propto \exp \{ \mathbb{E}_{p(\theta|z)} [\log p(x|z, \theta)] \} \quad (136)$$

$$\propto \exp \{ \langle t_z(x), \mathbb{E}_{p(\theta|z)} [\theta] \rangle \} \quad (137)$$

839 And so we can recognize that  $q^*(x|z)$  is in the same exponential family as  $p(x|z, \theta)$  but with natural  
 840 parameter  $\mathbb{E}_{p(\theta|z)} [\theta]$ . This completes the proof.  $\square$

841 Next, we emphasize another form of the CMFVI solution that is convenient when deriving CMFVI  
 842 solutions of other models.

843 **Lemma 5** (Mean field form of CMFVI solution). *The CMFVI approximation of  $p(x|z)$  has the*  
 844 *following form:*

$$q^*(x|z) \propto \exp \{ \mathbb{E}_{p(\theta|z)} [\log p(x|z, \theta)] \} \quad (138)$$

845 *Proof.* See Eq. [\(136\)](#)  $\square$

846 **Corollary 6** (Value of CMFVI objective at optimum). *The value of the CMFVI objective at the*  
 847 *optimum is given by:*

$$\text{KL} [q^*(x|z)p(\theta|z) || p(x, \theta|z)] = \mathbb{E}_{p(\theta|z)} [A(z, \theta)] - A(z, \theta^*(z)) \quad (139)$$

848 where  $z$  is fixed,  $\theta^*(z) = \mathbb{E}_{p(\theta|z)} [\theta]$  and  $A(z, \theta)$  is the partition function of  $p(x|z, \theta)$ .

849 *Proof.* Let  $\theta^*(z) = \mathbb{E}_{p(\theta|z)}[\theta]$ . Recall that  $p(x|z, \theta) = \exp\{\langle t_z(x), \theta \rangle - A(z, \theta)\}$ ,  $q^*(x|z) =$   
 850  $p(x|z, \theta^*(z))$  and that the CMFVI objective can be written using an identity from Lemma 4:

$$\text{KL}[q(x|z)p(\theta|z)||p(x, \theta|z)] = \mathbb{E}_{q(x|z)}[\log q(x|z) - \mathbb{E}_{p(\theta|z)}[\log p(x|z, \theta)]] \quad (140)$$

851 We can plug  $q^*(x|z)$  and  $p(x|z, \theta)$  into the identity to get:

$$\text{KL}[q^*(x|z)p(\theta|z)||p(x, \theta|z)] \quad (141)$$

$$= \mathbb{E}_{q^*(x|z)}[\log q^*(x|z) - \mathbb{E}_{p(\theta|z)}[\log p(x|z, \theta)]] \quad (142)$$

$$= \mathbb{E}_{q^*(x|z)} \left[ \left( \langle t_z(x), \theta^*(z) \rangle - A(z, \theta^*(z)) \right) - \left( \langle t_z(x), \underbrace{\mathbb{E}_{p(\theta|z)}[\theta]}_{\theta^*(z)} \rangle - \mathbb{E}_{p(\theta|z)}[A(z, \theta)] \right) \right] \quad (143)$$

$$= \mathbb{E}_{p(\theta|z)}[A(z, \theta)] - A(z, \theta^*(z)) \quad (144)$$

852 □

853 **Proposition 14** (Forward KL divergence). *The forward KL divergence between  $p(x|z)$  and  $q^*(x|z)$*   
 854 *is given by:*

$$\text{KL}[p(x|z)||q^*(x|z)] = -H_p[x|z] - \langle t^*(z), \theta^*(z) \rangle + A(z, \theta^*(z)) \quad (145)$$

855 where  $H_p[x|z]$  is the differential entropy of  $p(x|z)$ ,  $t^*(z) = \mathbb{E}_{p(x|z)}[t_z(x)]$ ,  $\theta^*(z) = \mathbb{E}_{p(\theta|z)}[\theta]$  and  
 856  $A(z, \theta)$  is the partition function of  $p(x|z, \theta)$ .

857 *Proof.* This follows from a direct computation:

$$\text{KL}[p(x|z)||q^*(x|z)] = -H_p[x|z] - \int p(x|z) \log q^*(x|z) dx \quad (146)$$

$$= -H_p[x|z] - \int p(x|z) (\langle t_z(x), \theta^*(z) \rangle - A(z, \theta^*(z))) dx \quad (147)$$

$$= -H_p[x|z] - \langle \int p(x|z) t_z(x) dx, \theta^*(z) \rangle + A(z, \theta^*(z)) \quad (148)$$

$$= -H_p[x|z] - \langle t^*(z), \theta^*(z) \rangle + A(z, \theta^*(z)) \quad (149)$$

858 □

## 859 F.2 Bayes estimator equivariance

860 We will use the equivariance of the Bayes estimator to linear transformations to show that it is also  
 861 equivariant to message passing updates when the Gaussian potential functions of the corresponding  
 862 CRF have covariances that only depend on the node index. This result will allow us to reparameterize  
 863 the Bayes estimator of the backward messages in terms of the previously computed backward  
 864 messages, and also in terms of the potential function means themselves. This will be useful for  
 865 relating the CMFVI time series models we construct back traditional time series models, and also  
 866 for proving that the autoregressive CMFVI model we construct is an approximation of flow-based  
 867 generative models for time series.

868 **Corollary 7** (Commutativity of Bayes estimator with update and marginalize opera-  
 869 tor). *Let  $\phi_{k+1|k}(x_{k+1}|x_k)$  be a Gaussian transition function and let  $\phi(x_{k+1}|\eta_{k+1}) :=$*   
 870  *$N(x_{k+1}|\mu_{k+1}(y), J_{k+1}^{-1})$  be a Gaussian node potential where  $y \sim p(y)$  is an auxiliary variable*  
 871 *set of variables that only the mean of the potential depends on. Then the Bayes estimator of  $\eta_k$*   
 872 *commutes with the update and marginalize operator. That is,*

$$\mathbb{E}_{p(y)}[\eta_k(y)] = \mathbb{E}_{p(y)}[\Phi_{k,k+1}(\eta_{k+1}(y))] = \Phi_{k,k+1}(\mathbb{E}_{p(y)}[\eta_{k+1}(y)]) \quad (150)$$

873 *Proof.* We can examine the form of  $\Phi_{k,k+1}$  from Corollary 2 to see that  $\Phi_{k,k+1}$  is linear with respect  
 874 to  $\mu_{k+1}(y)$ . Then the result follows from linearity equivariance of the Bayes estimator. □

### 875 F.3 CMFVI time series models

876 **Proposition 15** (Naive CMFVI solution). *Let  $p(x_{t_{1:N}}|y_{\mathcal{O}})$  be the target distribution. Then the naive*  
 877 *CMFVI solution, denoted by  $q^{CRF}(x_{t_{1:N}})$  is the CMFVI approximation of  $p(x_{t_{1:N}}|y_{\mathcal{O}})$  and is given*  
 878 *by:*

$$q^{CRF}(x_{t_{1:N}}) \propto \prod_{t_k \in \mathcal{T}} \phi_{t_{k+1}|t_k}(x_{t_{k+1}}|x_{t_k}) \prod_{t_k \in \mathcal{R}} \phi(x_{t_k}|\theta_{t_k}^*(y_{\mathcal{O}})) \quad (151)$$

879 where  $\theta_{t_k}^*(y_{\mathcal{O}}) = \mathbb{E}_{p(y_{\mathcal{U}}|y_{\mathcal{O}})}[\theta_{t_k}(y_{\tau_{1:T}})]$  is the Bayes estimator of  $\theta_{t_k}$ .

880 *Proof.* By expanding  $q^*$  using Lemma 5 one finds that the terms of the log likelihood is linear with  
 881 respect to  $\theta_{t_k}(y_{\tau_{1:T}})$ . Then the result follows from the equivariance of the Bayes estimator to linear  
 882 transformations.  $\square$

883 **Proposition 16** (CMFVI transition approximation). *Let  $p(x_{t_{1:N}}|y_{\mathcal{O}})$  be the target distribution and*  
 884 *consider its  $k$ 'th autoregressive factor  $p(x_{t_k}|x_{t_{1:k-1}}, y_{\mathcal{O}})$ . Then the CMFVI transition approximation*  
 885 *is given by:*

$$q^{transition}(x_{t_k}|x_{t_{1:k-1}}, y_{\mathcal{O}}) \propto \phi_{t_k|t_{k-1}}(x_{t_k}|x_{t_{k-1}})\phi(x_{t_k}|\beta_{t_k}^*(x_{t_{1:k-1}}, y_{\mathcal{O}})) \quad (152)$$

886 where  $\beta_{t_k}^*(x_{t_{1:k-1}}, y_{\mathcal{O}}) = \mathbb{E}_{p(y_{\mathcal{U}}|x_{t_{1:k-1}}, y_{\mathcal{O}})}[\beta_{t_k}(y_{\tau_{1:T}})]$  is the Bayes estimate of  $\beta_{t_k}(y_{\tau_{1:T}})$ , which is  
 887 defined using the message passing update operator  $\Phi_{t_k, t_{k+1}}$  from Definition 7 as:

$$\beta_{t_k} = \begin{cases} \Phi_{t_k, t_{k+1}}(\beta_{t_{k+1}}(y_{\tau_{1:T}}) + \theta_{t_{k+1}}(y_{\tau_{1:T}})) & \text{if } t_{k+1} \in \mathcal{R} \\ \Phi_{t_k, t_{k+1}}(\beta_{t_{k+1}}(y_{\tau_{1:T}})) & \text{otherwise} \end{cases} \quad (153)$$

888 *Proof.* The transition distribution in the fully observed setting is given by:

$$p(x_{t_k}|x_{t_{1:k-1}}, y_{\tau_{1:T}}) = p(x_{t_k}|x_{t_{k-1}}, y_{\tau_{1:T}}) \quad (154)$$

$$\propto \phi_{t_k|t_{k-1}}(x_{t_k}|x_{t_{k-1}})\phi(x_{t_k}|\beta_{t_k}(y_{\tau_{1:T}})) \quad (155)$$

889 If we expand the log likelihood of  $p(x_{t_k}|x_{t_{1:k-1}}, y_{\tau_{1:T}})$ , we would find that the log likelihood is linear  
 890 with respect to  $\beta_{t_k}(y_{\tau_{1:T}})$ , and so writing the CMFVI solution using Eq. (136) yields the result.  $\square$

891 We denote this model by  $q^{MSE}(x_{t_{1:N}}|y_{\mathcal{O}})$ .

892 **Corollary 8** (MSE Forecaster). *Let  $p(x_{t_{1:N}}|y_{\mathcal{O}})$  be the target distribution and suppose the co-*  
 893 *variances of its potentials are constant with respect to  $y$ . Then the MSE-CMFVI solution, de-*  
 894 *noted by  $q^{MSE}(x_{t_{1:N}})$  is the CMFVI approximation of  $p(x_{t_{1:N}}|y_{\mathcal{O}})$  obtained by choosing  $(x, z, \theta) =$*   
 895  *$(x_{t_{1:N}}, y_{\mathcal{O}}, \theta(y_{\tau_{1:T}}))$ :*

$$q^{MSE}(x_{t_{1:N}}|y_{\mathcal{O}}) \propto \prod_{t_k \in \mathcal{T}} \phi_{t_{k+1}|t_k}(x_{t_{k+1}}|x_{t_k}) \prod_{t_k \in \mathcal{R}} N(x_{t_k}|\mu_{t_k}^*(y_{\mathcal{O}}), \Sigma_{t_k}) \quad (156)$$

896 where  $\mu_{t_k}^*(y_{\mathcal{O}}) = \mathbb{E}_{p(y_{\mathcal{U}}|y_{\mathcal{O}})}[\mu_{t_k}(y_{\tau_{1:T}})]$  is the Bayes estimate of  $\mu_{t_k}$ , and  $\phi(x_{t_k}|\theta_{t_k}(y_{\tau_{1:T}})) =$   
 897  $N(x_{t_k}|\mu_{t_k}^*(y_{\tau_{1:T}}), \Sigma_{t_k})$ .

898 See Appendix F.3 for a proof.

899 **Definition 8** (Autoregressive CMFVI solution). *Let  $p(x_{t_{1:N}}|y_{\mathcal{O}})$  be the target distribution. Then the*  
 900 *autoregressive CMFVI solution, denoted by  $q^{AR}(x_{t_{1:N}})$  is the CMFVI approximation of  $p(x_{t_{1:N}}|y_{\mathcal{O}})$*   
 901 *and is given by:*

$$q^{AR}(x_{t_{1:N}}) \propto p(x_{t_1}|y_{\mathcal{O}}) \prod_{t_k \in \mathcal{T}} q^{transition}(x_{t_k}|x_{t_{1:k-1}}, y_{\mathcal{O}}) \quad (157)$$

902 where  $q^{transition}(x_{t_k}|x_{t_{1:k-1}}, y_{\mathcal{O}})$  is the CMFVI transition approximation given by Proposition 6

903 **Corollary 9** (MSE Forecaster). *Let  $p(x_{t_{1:N}}|y_{\mathcal{O}})$  be the target distribution and suppose the covari-*  
 904 *ances of its potentials are constant with respect to  $y$ . Then the MSE-CMFVI solution, denoted by*  
 905  *$q^{MSE}(x_{t_{1:N}})$  is the CMFVI approximation of  $p(x_{t_{1:N}}|y_{\mathcal{O}})$  and is given by:*

$$q^{MSE}(x_{t_{1:N}}) \propto \prod_{t_k \in \mathcal{T}} \phi_{t_{k+1}|t_k}(x_{t_{k+1}}|x_{t_k}) \prod_{t_k \in \mathcal{R}} N(x_{t_k}|\mu_{t_k}^*(y_{\mathcal{O}}), \Sigma_{t_k}) \quad (158)$$

906 where  $\mu_{t_k}^*(y_{\mathcal{O}}) = \mathbb{E}_{p(y_{\mathcal{U}}|y_{\mathcal{O}})}[\mu_{t_k}(y_{\tau_{1:T}})]$  is the Bayes estimate of  $\mu_{t_k}$ .

907 *Proof.* This follows from the fact that the potentials are constant with respect to  $y$  and the linear  
 908 equivariance of the Bayes estimator.  $\square$

909 **Corollary 10** (Autoregressive MSE Forecaster). *Let  $p(x_{t_{1:N}}|y_{\mathcal{O}})$  be the target distribution and*  
 910 *suppose the covariances of its potentials are constant with respect to  $y$ . Then the autoregressive*  
 911 *MSE-CMFVI solution, denoted by  $q^{\text{AR-MSE}}(x_{t_{1:N}})$  is the CMFVI approximation of  $p(x_{t_{1:N}}|y_{\mathcal{O}})$  and is*  
 912 *given by:*

$$q^{\text{AR-MSE}}(x_{t_{1:N}}) \propto p(x_{t_1}|y_{\mathcal{O}}) \prod_{t_k \in \mathcal{T}} \phi_{t_k|t_{k-1}}(x_{t_k}|x_{t_{k-1}}) \prod_{t_k \in \mathcal{R}} N(x_{t_k} | (\mu_{t_k}^{\beta})^*(x_{t_{1:k}}, y_{\mathcal{O}}), \Sigma_{t_k}^{\beta}) \quad (159)$$

913 where  $(\mu_{t_k}^{\beta})^*(x_{t_{1:k}}, y_{\mathcal{O}}) = \mathbb{E}_{p(y_{\mathcal{U}}|x_{t_{1:k}}, y_{\mathcal{O}})} [\mu_{t_k}^{\beta}(y_{\tau_{1:T}})]$  is the Bayes estimate of  $\mu_{t_k}^{\beta}$  and  $\Sigma_{t_k}^{\beta}$  is the  
 914 covariance of the backward message of  $p(x_{t_{1:N}}|y_{\tau_{1:T}})$ .

915 *Proof.* This follows from the fact that the potentials are constant with respect to  $y$  and the linear  
 916 equivariance of the Bayes estimator.  $\square$

917 **Definition 9** (Continuous extension of AR-MSE model). *Let  $q^{\text{AR}}$  be the autoregressive CMFVI*  
 918 *solution and consider the setting where the potential functions of  $p(x_{t_{1:N}}|y_{\tau_{1:T}})$  have covariances*  
 919 *that do not depend on  $y$ . Then the continuous extension of  $q^{\text{AR}}$  is given by the following piecewise*  
 920 *linear SDE:*

$$dx_t = (F_t x_t + L_t L_t^T \nabla \log \phi(x_t | \beta_t^*(x_{t_{1:k}}, y_{\mathcal{O}}))) dt + L_t dW_t, \quad (160)$$

$$\text{where } \beta_t^*(x_{t_{1:k}}, y_{\mathcal{O}}) = \mathbb{E}_{p(y_{\mathcal{U}}|x_{t_{1:k}}, y_{\mathcal{O}})} [\beta_t(y_{\tau_{1:T}})], \text{ and } t \in (t_k, t_{k+1}) \quad (161)$$

921 where  $\beta_t^*(x_{t_{1:k}}, y_{\mathcal{O}})$  is the Bayes estimator of  $\beta_t(y_{\tau_{1:T}}) = \Phi_{t, t_{k+1}}(\beta_{t_{k+1}}(y_{\tau_{1:T}}))$ .

922 *Proof.* We just need to verify that this piecewise linear SDE has the same joint distribution as  $q^{\text{AR}}$   
 923 on  $t_{1:N}$ . To do this, we can just check that each of the linear SDEs that are defined on the intervals  
 924  $(t_k, t_{k+1})$  have the same joint distribution as  $q^{\text{transition}}(x_{t_k}|x_{t_{1:k-1}}, y_{\mathcal{O}})$  from Proposition 6  $\square$

## 925 G Flow-based generative models proofs

926 In this section we provide basic results about Bayes estimation for generalized linear stochastic  
 927 interpolants. Let  $dx_t = F_t x_t dt + L_t dW_t$  be the base linear SDE and let the distribution of random  
 928 draws, at times  $t_{1:N}$ , be denoted by  $p(x_{t_{1:N}}|c)$ . Let  $p(x_{t_{1:N}}|\theta, c)$  be its conditional distribution given  
 929 parameters  $\theta$  that are only available during training time and some extra conditioning information  $c$   
 930 that is available at both training and test time, and suppose that  $p(\theta|c)$  is the (unknown) distribution of  
 931  $\theta$  given  $c$ . The goal of the techniques in this section (and FBGMs in general), is to construct, and  
 932 learn, the distribution of  $p(x_{t_{1:N}}|c)$ , which is the distribution needed to generate samples of  $x_{t_{1:N}}$   
 933 when we do not have access to the parameters  $\theta$ . At a high level, FBGMs offer different inference  
 934 algorithms for this task. In this section, we will derive three of these inference algorithms.

### 935 G.1 Score function for FBGMs

936 **Proposition 17** (Score function for FBGMs). *Suppose that  $p(\theta|c)$  is a probability distribution*  
 937 *over  $\theta$  given some extra conditioning information  $c$  and  $p(x_t|\theta, c)$  is the marginal distribution of a*  
 938 *generalized linear stochastic interpolant whose base linear SDE is given by  $dx_t = F_t x_t dt + L_t dW_t$ .*  
 939 *Then the score function of  $p(x_t|c)$  is given by:*

$$\nabla \log p(x_t|c) = \nabla \log \phi(x_t | \alpha_t^*(x_t, \theta, c) + \beta_t^*(x_t, \theta, c)) \quad (162)$$

940 where  $\alpha_t^*(x_t, \theta, c) = \mathbb{E}_{p(\theta|x_t, c)} [\alpha_t(\theta, c)]$  and  $\beta_t^*(x_t, \theta, c) = \mathbb{E}_{p(\theta|x_t, c)} [\beta_t(\theta, c)]$  are Bayes estimators  
 941 of the forward and backward messages to time  $t$  using  $x_t$  respectively.

942 *Proof.* A straightforward calculation will lead to the desired result.

$$\nabla \log p(x_t|c) = \frac{1}{p(x_t|c)} \nabla p(x_t|c) \quad (163)$$



$$= \frac{1}{p(x_t|c)} \nabla \int p(\theta|c) p(x_t|\theta, c) d\theta \quad (164)$$

$$= \frac{1}{p(x_t|c)} \int p(\theta|c) \nabla p(x_t|\theta, c) d\theta \quad (165)$$

$$= \int \frac{p(\theta|c) p(x_t|\theta, c)}{p(x_t|c)} \nabla \log p(x_t|\theta, c) d\theta \quad (166)$$

$$= \mathbb{E}_{p(\theta|x_t, c)} [\nabla \log p(x_t|\theta, c)] \quad (167)$$

$$= \mathbb{E}_{p(\theta|x_t, c)} [\nabla \log \phi(x_t|\alpha_t(\theta, c) + \beta_t(\theta, c))] \quad \because \text{Lemma 2} \quad (168)$$

$$= \nabla \log \phi(x_t|\alpha_t^*(x_t, \theta, c) + \beta_t^*(x_t, \theta, c)) \quad \because \text{Eq. (21)} \quad (169)$$

943

□

## 944 G.2 General form of Markovian projection SDE

945 **Lemma 6** (General form of Markovian projection SDE). *Suppose that  $p(\theta|c)$  is a probability*  
 946 *distribution over  $\theta$  given some extra conditioning information  $c$  and  $p(x_t|\theta, c)$  is the marginal*  
 947 *distribution of a generalized linear stochastic interpolant whose base linear SDE is given by  $dx_t =$*   
 948  *$F_t x_t dt + L_t dW_t$ . Then the Markovian projection SDE is given by:*

$$dx_t = (F_t x_t + L_t L_t^T \nabla \log \phi(x_t|\beta_t^*(x_t, \theta, c))) dt + L_t dW_t \quad (170)$$

949 where  $\beta_t^*(x_t, \theta, c) = \mathbb{E}_{p(\theta|x_t, c)} [\beta_t(\theta, c)]$  is the Bayes estimate of the backward message to time  $t$   
 950 using  $x_t$ .

951 *Proof.* The Markovian projection SDE is the SDE whose marginal distribution evolves in time in  
 952 the same way that  $p(x_t|c)$  evolves in time, and so our proof strategy will follow the same strategy  
 953 as [Lipman et al. 2023] Theorem 1] where we take the time derivative of  $p(x_t|c)$  and recognize the  
 954 form of the SDE.

955 First, recall that the Fokker-Planck equation [Särkkä and Solin 2019, Øksendal and Øksendal 2003]  
 956 relates an SDE to the time derivative of its marginal distribution. Let  $p(x_t|\theta, c)$  be the marginal  
 957 distribution of the generalized linear stochastic interpolant and recall that its corresponding SDE  
 958 is given by  $dx_t = (F_t x_t + L_t L_t^T \nabla \log \phi(x_t|\beta_t(\theta, c))) dt + L_t dW_t$  (see Proposition 4). Then the  
 959 Fokker-Planck equation for this SDE is given by:

$$\frac{\partial p(x_t|\theta, c)}{\partial t} = -\text{Div}(p(x_t|\theta, c)(F_t x_t + L_t L_t^T \nabla \log \phi(x_t|\beta_t(\theta, c)))) + \frac{1}{2} L_t L_t^T \text{Div}(\nabla p(x_t|\theta, c)) \quad (171)$$

960  $L_t L_t^T$  appears outside the divergence operator because it does not depend on  $x_t$ . Next, we can directly  
 961 take the time derivative of  $p(x_t|c)$  and recognize the form of the corresponding SDE.

$$\frac{\partial p(x_t|c)}{\partial t} = \mathbb{E}_{p(\theta|c)} \left[ \frac{\partial p(x_t|\theta, c)}{\partial t} \right] \quad (172)$$

$$= \mathbb{E}_{p(\theta|c)} \left[ -\text{Div}(p(x_t|\theta, c)(F_t x_t + L_t L_t^T \nabla \log \phi(x_t|\beta_t(\theta, c)))) + \frac{1}{2} L_t L_t^T \text{Div}(\nabla p(x_t|\theta, c)) \right] \quad (173)$$

$$= \mathbb{E}_{p(\theta|c)} [-\text{Div}(p(x_t|\theta, c) F_t x_t)] \quad (\text{A}) \quad (174)$$

$$+ \mathbb{E}_{p(\theta|c)} [-\text{Div}(p(x_t|\theta, c) L_t L_t^T \nabla \log \phi(x_t|\beta_t(\theta, c)))] \quad (\text{B}) \quad (175)$$

$$+ \mathbb{E}_{p(\theta|c)} \left[ \frac{1}{2} L_t L_t^T \text{Div}(\nabla p(x_t|\theta, c)) \right] \quad (\text{C}) \quad (176)$$

962 Since all of the divergence and gradient operators depend only on  $x_t$ , we can pass the expectation  
 963 through these terms. We can simplify each terms as follows:

(A)

$$\mathbb{E}_{p(\theta|c)} [-\text{Div}(p(x_t|\theta, c) F_t x_t)] = -\text{Div}(p(x_t|c) F_t x_t) \quad (177)$$

(B)

$$\mathbb{E}_{p(\theta|c)} [-\text{Div}(p(x_t|\theta, c)L_t L_t^T \nabla \log \phi(x_t|\beta_t(\theta, c)))] = -\text{Div} \left( \int p(\theta|c)p(x_t|\theta, c)L_t L_t^T \nabla \log \phi(x_t|\beta_t(\theta, c))d\theta \right) \quad (178)$$

$$= -\text{Div} \left( \int p(\theta|x_t, c)p(x_t|c)L_t L_t^T \nabla \log \phi(x_t|\beta_t(\theta, c))d\theta \right) \quad (179)$$

$$= -\text{Div}(p(x_t|c)L_t L_t^T \mathbb{E}_{p(\theta|x_t, c)} [\nabla \log \phi(x_t|\beta_t(\theta, c))]) \quad (180)$$

(C)

$$\mathbb{E}_{p(\theta|c)} \left[ \frac{1}{2} L_t L_t^T \text{Div}(\nabla p(x_t|\theta, c)) \right] = \frac{1}{2} L_t L_t^T \text{Div}(\nabla \mathbb{E}_{p(\theta|c)} [p(x_t|\theta, c)]) \quad (181)$$

$$= \frac{1}{2} L_t L_t^T \text{Div}(\nabla p(x_t|c)) \quad (182)$$

964 Putting these terms back together, we get:

$$\frac{\partial p(x_t|c)}{\partial t} = -\text{Div}(p(x_t|c) \underbrace{(F_t x_t + L_t L_t^T \mathbb{E}_{p(\theta|x_t, c)} [\nabla \log \phi(x_t|\beta_t(\theta, c))])}_{\text{recognize as drift term in Fokker-Planck equation}}) + \frac{1}{2} L_t L_t^T \text{Div}(\nabla p(x_t|c)) \quad (183)$$

965 We can see that the form of the Markovian projection SDE is given by:

$$dx_t = (F_t x_t + L_t L_t^T \mathbb{E}_{p(\theta|x_t, c)} [\nabla \log \phi(x_t|\beta_t(\theta, c))]) dt + L_t dW_t \quad (184)$$

966 Lastly because  $\phi(x_t|\beta_t(\theta, c))$  is a Gaussian distribution with natural parameters  $\beta_t(\theta, c)$ , its pdf is  
967 given by:

$$\phi(x_t|\beta_t(\theta, c)) = \exp\{\langle t_c(x_t), \beta_t(\theta, c) \rangle - A(c, \theta)\} \quad (185)$$

$$(186)$$

968 where  $t_c(x_t)$  is the sufficient statistic of the Gaussian distribution and  $A(c, \theta)$  is the log partition  
969 function. From this form, we can immediately see that the expectation around the score function  
970 passes through to the natural parameters:

$$\mathbb{E}_{p(\theta|x_t, c)} [\nabla \log \phi(x_t|\beta_t(\theta, c))] = \langle \nabla t_c(x_t), \mathbb{E}_{p(\theta|x_t, c)} [\beta_t(\theta, c)] \rangle \quad (187)$$

971 If we let  $\beta_t^*(x_t, \theta, c) = \mathbb{E}_{p(\theta|x_t, c)} [\beta_t(\theta, c)]$  and stop the gradient with respect to  $x_t$  through  $\beta_t^*$ , then  
972 we recover the desired result.  $\square$

973 **Proposition 18** (Neural latent SDE). *Let  $p(x_{1:N}, y_{1:T})$  be the joint distribution defined in Definition 2*  
974 *and suppose that  $\mathbf{y} = (y_{\mathcal{O}}, y_{\mathcal{U}})$ , where  $\mathcal{O}$  and  $\mathcal{U}$  are the times at which sequences are observed and*  
975 *unobserved, respectively. Then the neural latent SDE is the following piecewise SDE defined on the*  
976 *intervals  $(t_k, t_{k+1})$  for  $k = 1, \dots, N$ :*

$$dx_t = (F_t x_t + L_t L_t^T \nabla \log \phi(x_t|\beta_t^*(x_t, x_{t_{1:k}}, y_{\mathcal{O}})))dt + L_t dW_t, \quad (188)$$

$$\text{where } \beta_t^*(x_t, x_{t_{1:k}}, y_{\mathcal{O}}) = \mathbb{E}_{p(y_{\mathcal{U}}|x_t, x_{t_{1:k}}, y_{\mathcal{O}})} [\beta_t(y_{1:T})], \text{ and } t \in (t_k, t_{k+1}) \quad (189)$$

977  $\beta_t^*(x_t, x_{t_{1:k}}, y_{\mathcal{O}})$  is the Bayes estimator of  $\beta_t$  using the current state  $x_t$ .

978 *Proof.* The result follows directly from Lemma 6 by choosing  $\theta = y_{\mathcal{U}}$  and  $c = x_{t_{1:k}}$ .  $\square$

### 979 G.3 General form of Markovian projection ODE

980 **Lemma 7** (General form of Markovian projection ODE). *Suppose that  $p(\theta|c)$  is a probability*  
981 *distribution over  $\theta$  given some extra conditioning information  $c$  and  $p(x_t|\theta, c)$  is the marginal*  
982 *distribution of a generalized linear stochastic interpolant whose base linear SDE is given by  $dx_t =$*   
983  *$F_t x_t dt + L_t dW_t$ . Then the Markovian projection ODE is defined as the probability flow ODE of the*  
984 *Markovian projection SDE and is given by:*

$$\frac{dx_t}{dt} = F_t x_t + \frac{1}{2} L_t L_t^T (\nabla \log \phi(x_t|\beta_t^*(x_t, \theta, c)) - \nabla \log \phi(x_t|\alpha_t^*(x_t, \theta, c))) \quad (190)$$

985 where  $\beta_t^*(x_t, \theta, c) = \mathbb{E}_{p(\theta|x_t, c)} [\beta_t(\theta, c)]$  and  $\alpha_t^*(x_t, \theta, c) = \mathbb{E}_{p(\theta|x_t, c)} [\alpha_t(\theta, c)]$  are Bayes estimators  
986 of the forward and backward messages to time  $t$  using  $x_t$  respectively.

987 *Proof.* Recall that the definition of the probability flow ODE of an SDE of the form  $dx_t = u_t(x_t)dt +$   
 988  $L_t dW_t$  is given by [Song et al., 2021]:

$$\frac{dx_t}{dt} = u_t(x_t) - \frac{1}{2} L_t L_t^T \nabla \log p(x_t|c) \quad (191)$$

989 Plugging in drift of the Markovian projection SDE in Lemma 6 and the score function of  $p(x_t|c)$  in  
 990 Proposition 17 we get the desired result.  $\square$

## 991 H Message Passing Implementation Details

992 We devise a careful implementation of message passing to ensure numerical stability. There are many  
 993 different ways to implement message passing. For example, [Särkkä et al., 2006] parameterizes the  
 994 potentials in the standard form of Gaussians and uses Kalman filtering [Kalman, 1960] to obtain  
 995 the forward messages and does not directly compute the backward messages, but instead uses the  
 996 Rauch-Tung-Striebel smoother [Rauch et al., 1965] to blend the forward and backward message  
 997 computations to obtain the smoothed potentials. Alternatively, [Fox, 2009] [Johnson and Linderman,  
 998 2015] utilize a natural parameterization of the potentials in order to have simple message passing  
 999 updates. Our implementation requires that we can express both total uncertainty, and total certainty,  
 1000 in a variable in order to be able to work with incomplete, or missing data, and to condition exactly  
 1001 on variables. To do this, we adopt a mixed parametrization that contains the mean of the Gaussian  
 1002 and precision matrix so that we can express total uncertainty using a precision matrix of 0 and total  
 1003 certainty in the mean value by using a symbolic infinity. We also use symbolic zeros to mitigate  
 1004 accumulation of errors when perform message passing on long chains of latent variables without any  
 1005 evidence.

### 1006 H.1 Numerical stability considerations

1007 Before we look at the implementation details, we will look at what considerations we need to make  
 1008 for the implementation of these operations in a numerically stable way. Recall that the transition  
 1009 distribution of an LTI-SDE is given by

$$\phi(x_{t+s}|x_t) = N(x_{t+s}|A_s x_t, \Sigma_s) \quad (192)$$

1010 where

$$\begin{bmatrix} A_s & \Sigma_s A_s^{-T} \\ 0 & A_s^{-T} \end{bmatrix} := \exp\left\{ \begin{bmatrix} F & LL^T \\ 0 & -F^T \end{bmatrix} s \right\} \quad (193)$$

1011 and that potential functions can be written in natural or standard form as:

$$\phi(x) = \exp\left\{ -\frac{1}{2} x^T J x + x^T h - \log Z \right\} \quad (194)$$

$$= \exp\left\{ -\frac{1}{2} x^T \Sigma^{-1} x + x^T \Sigma^{-1} \mu - \log Z \right\} \quad (195)$$

1012 where  $\Sigma = J^{-1}$  and  $\mu = J^{-1}h$ . We assume that the time intervals between consecutive variables  
 1013 are bounded and nonzero so that  $\Sigma_s$ ,  $A_s$ , and  $A_s^{-T}$  are numerically stable. We also assume that the  
 1014 covariance matrices that the user specifies for the node potentials, e.g.  $\Sigma$  or  $J$ , are well conditioned.  
 1015 We do not assume that  $\Sigma_s^{-1}$ ,  $\Sigma^{-1}$  nor  $J^{-1}$  are well conditioned. These assumptions are made to  
 1016 accomodate operations that a user might perform in practice. For example, a user may choose to  
 1017 express 0 certainty in a variable by setting  $\Sigma \rightarrow \infty$  or  $J = 0$  and can choose to express 0 uncertainty  
 1018 by setting  $\Sigma = 0$  or  $J \rightarrow \infty$ . Furthermore, if a user chooses to discretize an SDE at points where  
 1019  $s$  is small, or even exactly 0, then  $\Sigma_s$  is close to 0 and so  $\Sigma_s^{-1}$  can be very large. To account  
 1020 for these considerations, we use symbolic computation to represent matrices that are 0 or  $\infty$  as  
 1021 needed. Furthermore, we use three different parameterizations of the Gaussian to ensure that we  
 1022 can handle all cases. We use the **standard** parameterization,  $(\mu, \Sigma)$ , **natural** parameterization<sup>3</sup>  
 1023  $(J = \Sigma^{-1}, h = \Sigma^{-1}\mu)$ , and **mixed** parameterization  $(J = \Sigma^{-1}, \mu)$ . For brevity, we will not include  
 1024 the updates for the normalizing constant  $\log Z$  in our pseudocode.

<sup>3</sup>The true natural parameters are scaled by  $-\frac{1}{2}$

## 1025 H.2 Message passing pseudocode

1026 In Appendix D we identified the key operations that are needed to perform variable elimination in the  
 1027 sequential and parallel settings (see Appendices D.1 and D.2). These operations are:

- 1028 1. An “add” operation adds the parameters of two potential functions together (code in Ap-  
 1029 pendix H.3).
- 1030 2. An “update” operation that absorbs a potential function into a transition function (defined in  
 1031 Definition 5 and code in Appendix H.3).
- 1032 3. A “marginalize” operation that marginalizes out a variable from a Gaussian joint distribution.  
 1033 In practice, we fuse this with the “update” operation (code in Appendix H.3).
- 1034 4. A “reverse” operation that reverses the direction of a transition (code in Appendix H.3).
- 1035 5. A “chain” operation that chains two transition functions (defined in Eq. (40) and code in  
 1036 Appendix H.3).

1037 In Appendix H.3, Appendix H.3, Appendix H.3, and Appendix H.3 we provide pseudocode for  
 1038 message passing that involves these operations.

## 1039 H.3 Update rules

Now we provide pseudocode for the update rules.

---

### Algorithm 1 Add

---

1. Require: potential functions  $\phi_1$  and  $\phi_2$
  2.  $(J_1, h_1) = \text{to\_natural}(\phi_1)$
  3.  $(J_2, h_2) = \text{to\_natural}(\phi_2)$
  4. Return  $\text{from\_natural}((J_1 + J_2, h_1 + h_2))$
- 

1040

---

### Algorithm 2 Update

---

1. Require: potential function  $\phi$  and transition  $\phi_{k+1|k}$
  2.  $(J, \mu) = \text{to\_mixed}(\phi)$
  3.  $(A, u, \Sigma) = \phi_{k+1|k}$
  4.  $R = J(I + \Sigma J)^{-1}$
  5.  $S = \Sigma R$
  6.  $T = I - S$
  7.  $\bar{\phi}_{k+1|k} = (TA, Tu + S\mu, T\Sigma)$
  8.  $\bar{\phi} = \text{from\_mixed}((A^T R^T A, A^{-1}(\mu - u)))$
  9.  $\Psi_{k+1,k} = (\bar{\phi}_{k+1|k}, \bar{\phi})$
  10. Return  $\Psi_{k+1,k}$
- 

---

### Algorithm 3 Update and marginalize

---

1. Require: potential function  $\phi$  and transition  $\phi_{k+1|k}$
  2.  $(\_, \bar{\phi}) = \text{Update}(\phi, \phi_{k+1|k})$
  3. Return  $\bar{\phi}$
-

---

**Algorithm 4 Reverse**

---

1. Require: transition  $\phi_{k+1|k}$
  2.  $(A, u, \Sigma) = \phi_{k+1|k}$
  3.  $\bar{A} = A^{-1}$
  4.  $\bar{u} = -A^{-1}u$
  5.  $\bar{\Sigma} = A^{-1}\Sigma A^{-T}$
  6. Return  $(\bar{A}, \bar{u}, \bar{\Sigma})$
- 

---

**Algorithm 5 Chain**

---

1. Require: transition functions  $\phi_{k|k-1}$  and  $\phi_{k+1|k}$
  2.  $A_k, u_k, \Sigma_k = \phi_{k+1|k}$
  3.  $A_{k-1}, u_{k-1}, \Sigma_{k-1} = \phi_{k|k-1}$
  4.  $A = A_k A_{k-1}$
  5.  $u = A_k u_{k-1} + u_k$
  6.  $\Sigma = \Sigma_k + A_k \Sigma_{k-1} A_k^T$
  7. Return  $(A, u, \Sigma)$
- 

---

**Algorithm 6 BackwardMessagePassing**

---

1. Require  $(\phi_{2|1}, \dots, \phi_{N|N-1})$  and  $(\phi_1, \dots, \phi_N)$
  2. Initialize  $\beta_N = 0$
  3. For  $k = N, \dots, 2$ :
    - (a)  $\Psi_{k,k-1} = \text{Update}(\phi_{k|k-1}, \phi_k + \beta_k)$
    - (b)  $\beta_{k-1} = \text{Marginalize}(\Psi_{k,k-1})$
  4. Return  $(\beta_1, \dots, \beta_N)$
- 

---

**Algorithm 7 ParallelBackwardMessagePassing**

---

1. Require  $(\phi_{2|1}, \dots, \phi_{N|N-1})$  and  $(\phi_1, \dots, \phi_N)$
  2. In parallel, for  $k = N, \dots, 2$ :
    - (a)  $\Psi_{k,k-1} = \text{Update}(\phi_{k|k-1}, \phi_k)$
  3.  $(\Psi_{1:N}, \dots, \Psi_{N-1:N}) = \text{AssociativeScan}(\text{Chain}, \Psi_{2,1}, \dots, \Psi_{N,N-1})$
  4. In parallel, for  $k = N-1, \dots, 1$ :
    - (a)  $\beta_k = \text{Marginalize}(\Psi_{k:N})$
  5.  $\beta_N = 0$
  6. Return  $(\beta_1, \dots, \beta_N)$
-

---

**Algorithm 8** ForwardMessagePassing

---

1. Require  $(\phi_{2|1}, \dots, \phi_{N|N-1}), (\phi_1, \dots, \phi_N)$  and `use_parallel`
  2. For  $k = 1, \dots, N - 1$ :
    - (a)  $\phi_{k|k+1} = \text{Reverse}(\phi_{k+1|k})$
  3. If `use_parallel`:
    - (a) `MessagePassing = ParallelBackwardMessagePassing`
  4. Else:
    - (a) `MessagePassing = BackwardMessagePassing`
  5.  $(\alpha_N, \dots, \alpha_1) = \text{MessagePassing}((\phi_{N-1|N}, \dots, \phi_{1|2}), (\phi_N, \dots, \phi_1))$
  6. Return  $(\alpha_1, \dots, \alpha_N)$
- 

---

**Algorithm 9** AssociativeScan (Even number of elements only)

---

1. Require: operator  $\oplus$ , elements  $(t_1, t_2, \dots, t_n)$  where  $n$  is a power of 2
  2. If  $n == 1$ :
    - (a) Return  $t_1$
  3. In parallel, for  $k = 1, \dots, n/2$ :
    - (a)  $p_k = t_{2k-1} \oplus t_{2k}$
  4.  $(r_2, r_4, \dots, r_n) = \text{AssociativeScan}(\oplus, (p_1, p_2, \dots, p_{n/2}))$
  5. In parallel, for  $k = 1, \dots, n/2 - 1$ :
    - (a)  $r_{2k+1} = r_{2k} \oplus t_{2k+1}$
  6.  $r_1 = t_1$
  7. Return  $(r_1, r_2, \dots, r_n)$
- 

## 1041 I Dataset details

1042 **Double pendulum** We constructed a synthetic task of forecasting the position and velocity of  
1043 a double pendulum from noisy observations of the position. We constructed a double pendulum  
1044 with two rods both with mass and length of 1 and 1 respectively and simulate its motion for 50,000  
1045 seconds and record the Euclidean space coordinates of the endpoints of each rod at a rate of 5Hz.  
1046 We then add Gaussian noise with a standard deviation of 0.3 to the position data and use the last  
1047 10000 points of the data as our full dataset. The conditioned linear SDE that we construct (to model  
1048  $p(\mathbf{x}|\mathbf{y}_{1:N})$ ) for this task is a Weiner velocity model with a diffusion coefficient of 4.0 to model the  
1049 latent space process and Gaussian potential functions centered at the noisy observed positions (with  
1050 zero padding for the velocity dimensions) and with a standard deviation of 0.3 over the position  
1051 dimensions and  $\infty$  over the velocity dimensions to ensure that the latent position coordinates can be  
1052 interpreted as smoothed versions of the observed positions and velocities.

1053 **Dynamical systems** The remaining datasets are similar to those defined in appendix A.1 of [El-  
1054 Gazzar and van Gerven, 2025]. The difference in our work is that we do not add process noise to  
1055 these dynamical systems during simulation and instead add varying amounts of Gaussian noise with  
1056 a to the trajectories that we generate. For Brusselator, Lotka and Van der Pol we add Gaussian noise  
1057 with a standard deviation of 0.3 and for FitzHugh and Lorenz we add noise with a standard deviation  
1058 of 0.2 and 1.0 respectively.

## 1059 J Model implementation details

### 1060 J.1 Base SDE

1061 In all of our experiments, we used the Wiener velocity model [Särkkä and Solin 2019] as our base  
1062 SDE. In this model, we assume that the latent variable  $x_t$  is the concatenation of a ground truth  
1063 position, denoted by  $z_t$ , and velocity, denoted by  $v_t$ . To capture the physical relationship between  
1064 position and velocity, we set the relationship  $\frac{dz_t}{dt} = v_t$  by constructing  $F$  to be the appropriate block  
1065 matrix. Finally, to ensure that the paths that we sample are smooth, we set  $L$  to be 0 in the entries  
1066 corresponding to  $z_t$ . This linear time-invariant SDE takes the following form

$$d \begin{bmatrix} z_t \\ v_t \end{bmatrix} = \begin{bmatrix} 0 & I \\ 0 & 0 \end{bmatrix} \begin{bmatrix} z_t \\ v_t \end{bmatrix} dt + \begin{bmatrix} 0 & 0 \\ 0 & \sigma I \end{bmatrix} dW_t \quad (196)$$

1067 For the Brusselator, double pendulum, Fitz Hugh, Lorenz, Lotka and Van der Pol datasets, we used a  
1068 value of  $\sigma$  equal to 0.1, 0.1, 0.379, 0.435, 0.1 and 0.1 respectively.

### 1069 J.2 Neural network architecture and training details

1070 We used an encoder/decoder neural network architecture to condition on the observed sequence  $y_{\tau_{1:T}}$ ,  
1071 and generate the latent sequence  $x_{t_{1:N}}$ . The encoder and decoder for all of our models were a single  
1072 layer GRU-RNN with a hidden layer size of 128 for all of our models. We initially used a more  
1073 complex transformer based architecture but found that the simple RNN worked as well in our simple  
1074 experimental setting. We incorporated information about the times in each series by constructing  
1075 a feature vector for each scalar time and concatenating it with the observed sequence of variables  
1076 before passing the concatenation to the RNN.

1077 Each of our models were trained on a single 2080ti GPU using a learning rate of  $10^{-4}$  using the  
1078 adamw optimizer, linear warmup of 1000 steps, and an effective batch size of 256 (we used a batch  
1079 size of 64 and 4 gradient accumulation steps). For each experiment, we used 5 random seeds to  
1080 initialize the model parameters and to split the data into training, validation, and test sets using an  
1081 80/10/10 split. We evaluated the objective function on the entire validation set every 1000 gradient  
1082 updates and stopped training when the value of the objective function over the entire validation set  
1083 stopped improving for 5 evaluations.

### 1084 J.3 Model details

1085 **MSE forecaster** The MSE forecaster predicts the mean of the potential functions of the CRF used  
1086 to construct the latent process. This model is trained to minimize the mean squared error between  
1087 the predicted mean of each potential function, and the mean of the potential function of the target  
1088 process. To generate samples from this model, we use the input  $y_{\tau_{1:k}}$  to generate the means of the  
1089 CRF potentials for the entire sequence of generated variables. We then sample from the CRF defined  
1090 by these potentials to get a sample from this model.

1091 **Autoregressive (AR) Models** The autoregressive models represent a conditional Gaussian chain  
1092 where the parameters of the next backward message are parametrized by a neural network (see  
1093 Proposition 7). To generate the next sample of a sequence, an input sequence  $y_{\mathcal{O}}$  and history  
1094 of generated latent values  $x_{t_{1:k}}$  are passed to a neural network that outputs the parameters of  
1095  $\phi(x_{t_{k+1}} | \beta_{t_{k+1}}^*(x_{t_{1:k}}, y_{\mathcal{O}}))$ . For the MSE models, we only parameterize the mean of this distribution  
1096 as in Proposition 7, and for the maximum likelihood models (MLE), we parametrize the mean  
1097 and diagonal covariance matrix. We then compute the transition distribution  $q(x_{t_{k+1}} | x_{t_{1:k}}, y_{\mathcal{O}}) \propto$   
1098  $\phi_{t_{k+1}|t_k}(x_{t_{k+1}} | x_{t_k}) \beta_{t_{k+1}}^*(x_{t_{1:k}}, y_{\mathcal{O}})$  and then draw the next element of the sequence from this distri-  
1099 bution. We compute  $q(x_{t_{k+1}} | x_{t_{1:k}}, y_{\mathcal{O}})$  by using the update operation in Appendix H.3

1100 **Diffusion model (FBGM)** The diffusion model is trained using flow-matching [Lipman et al.,  
1101 2023] using a brownian bridge between a Gaussian random variable and the sequence of unobserved  
1102 variables. This model is effectively the same as standard diffusion models for images, but applied  
1103 to a flattened time series vector. The decoder transformer network outputs the vector field of the  
1104 probability flow ODE that is used to simulate the process. Samples are generated by passing a  
1105 sequence of Gaussian random variables of the same size as  $y_{\tau_{k+1:N}}$  to an ODE solver that uses the  
1106 vector field output by the decoder to simulate the process.

## 1107 NeurIPS Paper Checklist

1108 The checklist is designed to encourage best practices for responsible machine learning research,  
1109 addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove  
1110 the checklist: **The papers not including the checklist will be desk rejected.** The checklist should  
1111 follow the references and follow the (optional) supplemental material. The checklist does NOT count  
1112 towards the page limit.

1113 Please read the checklist guidelines carefully for information on how to answer these questions. For  
1114 each question in the checklist:

- 1115 • You should answer [Yes], [No], or [NA].
- 1116 • [NA] means either that the question is Not Applicable for that particular paper or the  
1117 relevant information is Not Available.
- 1118 • Please provide a short (1-2 sentence) justification right after your answer (even for NA).

1119 **The checklist answers are an integral part of your paper submission.** They are visible to the  
1120 reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it  
1121 (after eventual revisions) with the final version of your paper, and its final version will be published  
1122 with the paper.

1123 The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation.  
1124 While "[Yes]" is generally preferable to "[No]", it is perfectly acceptable to answer "[No]" provided a  
1125 proper justification is given (e.g., "error bars are not reported because it would be too computationally  
1126 expensive" or "we were unable to find the license for the dataset we used"). In general, answering  
1127 "[No]" or "[NA]" is not grounds for rejection. While the questions are phrased in a binary way, we  
1128 acknowledge that the true answer is often more nuanced, so please just use your best judgment and  
1129 write a justification to elaborate. All supporting evidence can appear either in the main paper or the  
1130 supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification  
1131 please point to the section(s) where related material for the question can be found.

### 1132 1. Claims

1133 Question: Do the main claims made in the abstract and introduction accurately reflect the  
1134 paper's contributions and scope?

1135 Answer: [Yes]

1136 Justification: We introduced a generalization of the key elements of flow-based generative  
1137 models that are relevant to the time series setting and showed how this can be used to  
1138 construct related discrete time models.

1139 Guidelines:

- 1140 • The answer NA means that the abstract and introduction do not include the claims  
1141 made in the paper.
- 1142 • The abstract and/or introduction should clearly state the claims made, including the  
1143 contributions made in the paper and important assumptions and limitations. A No or  
1144 NA answer to this question will not be perceived well by the reviewers.
- 1145 • The claims made should match theoretical and experimental results, and reflect how  
1146 much the results can be expected to generalize to other settings.
- 1147 • It is fine to include aspirational goals as motivation as long as it is clear that these goals  
1148 are not attained by the paper.

### 1149 2. Limitations

1150 Question: Does the paper discuss the limitations of the work performed by the authors?

1151 Answer: [Yes]

1152 Justification: In section 3.4 and 3.6 we explained how the class of models we introduced are  
1153 ultimately just mean squared error based conditional Gaussian models and therefore may  
1154 not work as well in practice as their maximum likelihood counterparts on more stochastic  
1155 data.

1156 Guidelines:



- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We provide all of our proofs in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide all of our implementation details in the appendix and provide our code as supplementary material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.

- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We include our code as supplementary material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “NA” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

1265 Question: Does the paper specify all the training and test details (e.g., data splits, hyper-  
1266 parameters, how they were chosen, type of optimizer, etc.) necessary to understand the  
1267 results?

1268 Answer: [Yes]

1269 Justification: We explain our experimental setting in the experiments section

1270 Guidelines:

- 1271 • The answer NA means that the paper does not include experiments.
- 1272 • The experimental setting should be presented in the core of the paper to a level of detail  
1273 that is necessary to appreciate the results and make sense of them.
- 1274 • The full details can be provided either with the code, in appendix, or as supplemental  
1275 material.

## 1276 7. Experiment statistical significance

1277 Question: Does the paper report error bars suitably and correctly defined or other appropriate  
1278 information about the statistical significance of the experiments?

1279 Answer: [Yes]

1280 Justification: We provide the mean and standard error for the models trained in our experi-  
1281 ments.

1282 Guidelines:

- 1283 • The answer NA means that the paper does not include experiments.
- 1284 • The authors should answer "Yes" if the results are accompanied by error bars, confi-  
1285 dence intervals, or statistical significance tests, at least for the experiments that support  
1286 the main claims of the paper.
- 1287 • The factors of variability that the error bars are capturing should be clearly stated (for  
1288 example, train/test split, initialization, random drawing of some parameter, or overall  
1289 run with given experimental conditions).
- 1290 • The method for calculating the error bars should be explained (closed form formula,  
1291 call to a library function, bootstrap, etc.)
- 1292 • The assumptions made should be given (e.g., Normally distributed errors).
- 1293 • It should be clear whether the error bar is the standard deviation or the standard error  
1294 of the mean.
- 1295 • It is OK to report 1-sigma error bars, but one should state it. The authors should  
1296 preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis  
1297 of Normality of errors is not verified.
- 1298 • For asymmetric distributions, the authors should be careful not to show in tables or  
1299 figures symmetric error bars that would yield results that are out of range (e.g. negative  
1300 error rates).
- 1301 • If error bars are reported in tables or plots, The authors should explain in the text how  
1302 they were calculated and reference the corresponding figures or tables in the text.

## 1303 8. Experiments compute resources

1304 Question: For each experiment, does the paper provide sufficient information on the com-  
1305 puter resources (type of compute workers, memory, time of execution) needed to reproduce  
1306 the experiments?

1307 Answer: [Yes]

1308 Justification: We provide these details in the appendix.

1309 Guidelines:

- 1310 • The answer NA means that the paper does not include experiments.
- 1311 • The paper should indicate the type of compute workers CPU or GPU, internal cluster,  
1312 or cloud provider, including relevant memory and storage.
- 1313 • The paper should provide the amount of compute required for each of the individual  
1314 experimental runs as well as estimate the total compute.

- 1315           • The paper should disclose whether the full research project required more compute  
1316           than the experiments reported in the paper (e.g., preliminary or failed experiments that  
1317           didn't make it into the paper).

1318       **9. Code of ethics**

1319       Question: Does the research conducted in the paper conform, in every respect, with the  
1320       NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

1321       Answer: [Yes]

1322       Justification: We read the code of ethics.

1323       Guidelines:

- 1324           • The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.  
1325           • If the authors answer No, they should explain the special circumstances that require a  
1326           deviation from the Code of Ethics.  
1327           • The authors should make sure to preserve anonymity (e.g., if there is a special consid-  
1328           eration due to laws or regulations in their jurisdiction).

1329       **10. Broader impacts**

1330       Question: Does the paper discuss both potential positive societal impacts and negative  
1331       societal impacts of the work performed?

1332       Answer: [NA]

1333       Justification: Our paper is mostly theoretical with limited societal impacts at this stage.

1334       Guidelines:

- 1335           • The answer NA means that there is no societal impact of the work performed.  
1336           • If the authors answer NA or No, they should explain why their work has no societal  
1337           impact or why the paper does not address societal impact.  
1338           • Examples of negative societal impacts include potential malicious or unintended uses  
1339           (e.g., disinformation, generating fake profiles, surveillance), fairness considerations  
1340           (e.g., deployment of technologies that could make decisions that unfairly impact specific  
1341           groups), privacy considerations, and security considerations.  
1342           • The conference expects that many papers will be foundational research and not tied  
1343           to particular applications, let alone deployments. However, if there is a direct path to  
1344           any negative applications, the authors should point it out. For example, it is legitimate  
1345           to point out that an improvement in the quality of generative models could be used to  
1346           generate deepfakes for disinformation. On the other hand, it is not needed to point out  
1347           that a generic algorithm for optimizing neural networks could enable people to train  
1348           models that generate Deepfakes faster.  
1349           • The authors should consider possible harms that could arise when the technology is  
1350           being used as intended and functioning correctly, harms that could arise when the  
1351           technology is being used as intended but gives incorrect results, and harms following  
1352           from (intentional or unintentional) misuse of the technology.  
1353           • If there are negative societal impacts, the authors could also discuss possible mitigation  
1354           strategies (e.g., gated release of models, providing defenses in addition to attacks,  
1355           mechanisms for monitoring misuse, mechanisms to monitor how a system learns from  
1356           feedback over time, improving the efficiency and accessibility of ML).

1357       **11. Safeguards**

1358       Question: Does the paper describe safeguards that have been put in place for responsible  
1359       release of data or models that have a high risk for misuse (e.g., pretrained language models,  
1360       image generators, or scraped datasets)?

1361       Answer: [NA]

1362       Justification: Our method does not require safeguards.

1363       Guidelines:

- 1364           • The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: We wrote the code for our models and datasets from scratch.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

## 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: N/A

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

## 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: N/A

1416 Guidelines:

1417 • The answer NA means that the paper does not involve crowdsourcing nor research with

1418 human subjects.

1419 • Including this information in the supplemental material is fine, but if the main contribu-

1420 tion of the paper involves human subjects, then as much detail as possible should be

1421 included in the main paper.

1422 • According to the NeurIPS Code of Ethics, workers involved in data collection, curation,

1423 or other labor should be paid at least the minimum wage in the country of the data

1424 collector.

1425 **15. Institutional review board (IRB) approvals or equivalent for research with human**

1426 **subjects**

1427 Question: Does the paper describe potential risks incurred by study participants, whether

1428 such risks were disclosed to the subjects, and whether Institutional Review Board (IRB)

1429 approvals (or an equivalent approval/review based on the requirements of your country or

1430 institution) were obtained?

1431 Answer: [NA]

1432 Justification: N/A

1433 Guidelines:

1434 • The answer NA means that the paper does not involve crowdsourcing nor research with

1435 human subjects.

1436 • Depending on the country in which research is conducted, IRB approval (or equivalent)

1437 may be required for any human subjects research. If you obtained IRB approval, you

1438 should clearly state this in the paper.

1439 • We recognize that the procedures for this may vary significantly between institutions

1440 and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the

1441 guidelines for their institution.

1442 • For initial submissions, do not include any information that would break anonymity (if

1443 applicable), such as the institution conducting the review.

1444 **16. Declaration of LLM usage**

1445 Question: Does the paper describe the usage of LLMs if it is an important, original, or

1446 non-standard component of the core methods in this research? Note that if the LLM is used

1447 only for writing, editing, or formatting purposes and does not impact the core methodology,

1448 scientific rigorousness, or originality of the research, declaration is not required.

1449 Answer: [NA]

1450 Justification: We do not use LLMs in this work.

1451 Guidelines:

1452 • The answer NA means that the core method development in this research does not

1453 involve LLMs as any important, original, or non-standard components.

1454 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>)

1455 for what should or should not be described.