Amortized Structural Variational Inference

Shitao Fan University of Maryland
Carlos Misael Padilla Madrid Washington University in St. Louis
Yun Yang University of Maryland
Lizhen Lin University of Maryland

Abstract

Variational inference (VI) is widely used for approximate Bayesian inference, but it can scale poorly and often requires re-optimization when new data arrive. Amortized variational inference (AVI) learns a global inference map, yet standard mean-field AVI can suffer from large variational and amortization gaps because of independence assumptions. We propose amortized structural variational inference (ASVI), which injects structural dependencies among latent variables through neural architectures that encode local neighborhood information. ASVI reduces both gaps while retaining scalability. Simulations and real-data experiments show that ASVI improves predictive accuracy and posterior fidelity over AVI, and matches structured VI at lower computational cost.

1 INTRODUCTION

Variational inference (VI) (Blei et al., 2017; Wainwright et al., 2008) has become a widely used framework for approximate Bayesian inference, especially in high-dimensional and large-scale problems. By reformulating posterior inference as an optimization problem over a tractable family of distributions, VI offers a practical alternative to traditional sampling-based methods such as Markov chain Monte Carlo (MCMC), which are often computationally intensive. VI has been widely applied in statistical learning, probabilistic modeling, and deep generative models, most notably in learning encoder architectures within the variational autoencoder (VAE) framework (Kingma and Welling, 2013). Despite these benefits, standard VI methods can face scalability challenges with respect to sample size: as datasets grow larger, the per-data-point optimization of variational parameters becomes increasingly costly, both computationally and statistically. Moreover, the arrival of new data typically requires re-optimizing the variational objective from scratch, which limits the efficiency of VI in online or streaming data settings.

These limitations raise a natural question: can one design a variational inference framework whose per-instance computational cost remains stable as the dataset grows? A promising direction is *amortized variational inference* (AVI) (Ganguly et al., 2023; Zhang et al., 2018; Margossian and Blei, 2023; Agrawal and Domke, 2021), which draws on the idea of amortization from accounting, where a fixed cost is distributed across many units to reduce the per-unit burden. In *Bayesian latent variable models*, such as topic models (Blei et al., 2003; Jelodar et al., 2019), finite mixture models (Roberts et al., 1998; Nasios and Bors, 2006), state-space models (Geweke and Tanizaki, 2001; West and Harrison, 2006), and deep generative models like the VAE, each observation is typically associated with its own local latent variable, requiring a separate set of variational parameters. Standard VI therefore incurs a growing computational cost as data size increases. AVI addresses this by replacing the independent optimization of local variational parameters with a shared inference function that maps each observation to its corresponding variational parameters. This map is usually modeled by a deep neural network (DNN), trained jointly with the model to approximate the posterior over local latent variables. Once trained, this global inference network enables efficient and scalable posterior inference for each observation, with per-instance cost that does not increase with the overall data size.

While amortized variational inference (AVI) improves scalability, it introduces two key sources of approximation error: the variational gap and the amortization gap. The variational gap arises from the use of restricted variational families, such as the mean-field family, which are often too simplistic to capture dependencies among latent variables. The amortization gap refers to the discrepancy between the optimal variational parameters and those produced by the inference network. This gap is primarily due to limitations in the expressiveness or capacity of the learned inference function (Cremer et al., 2018). Recent efforts to reduce the amortization gap include the use of more expressive network architectures, regularization techniques, and hybrid

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

optimization methods that combine amortized inference with per-instance fine-tuning (e.g., (Zhang et al., 2022; Kim and Pavlovic, 2021; Krishnan et al., 2018)). However, many of these methods remain heuristic or introduce significant additional computational cost.

A parallel line of work has explored the construction of richer variational families to improve approximation fidelity, particularly in Bayesian models with dependent latent variables. For instance, Xing et al. (2012) introduced structured variational approximations to better capture dependencies among latent variables. Similarly, Wang et al. (2022) incorporated structural information into variational inference for latent state models, with a focus on the state-space model. Zhao and Linderman (2023) summarizes the idea of using structured ideas in the VAE setting. These advances suggest that enriching the structure of the variational family may offer a promising path for addressing the limitations of standard amortized methods, particularly those based on the mean-field approximating family, which assumes a fully factorized posterior and thus fails to capture dependencies among latent variables.

Building on these insights, we propose *amortized structural variational inference (ASVI)*, a novel framework that enhances amortized inference by constructing an amortization scheme over a structural variational family, in contrast to the standard factorized families commonly used in the AVI literature. ASVI integrates amortization with structured variational approximations, allowing the inference network to leverage model-induced structural dependencies among latent variables. Although prior work has incorporated structural components in specific settings, such as neighborhood-aware amortization in graph neural variational encoders Kipf and Welling (2016) and temporal VAEs with recurrent architectures Fraccaro et al. (2016), ASVI generalizes and formalizes these ideas within a unified variational framework. Rather than focusing exclusively on local interactions, it provides a flexible mechanism for embedding structural information into both the variational family and the inference network. As a result, ASVI offers a scalable and theoretically grounded approach that reduces both the variational and amortization gaps while retaining the computational efficiency of amortized inference.

The main contributions of this work are summarized as follows:

1. We introduce the ASVI framework, which incorporates structural information into amortized inference by jointly specifying a structured variational family and designing a structure-aware inference network. This integration reduces both the variational and amortization gaps, leading to improved approximation quality.
2. We provide theoretical guarantees for the ASVI framework by deriving explicit risk bounds for the resulting variational approximations. In particular, we analyze how architectural properties of the inference network, influence the approximation error and residual amortization gap. Our results demonstrate that incorporating local latent neighborhoods can substantially reduce the amortization gap while maintaining computational scalability.
3. We demonstrate the effectiveness of ASVI through extensive numerical experiments. Our results demonstrate that (a) ASVI maintains computational efficiency, (b) it significantly reduces the amortization gap, and (c) it achieves smaller variational gaps compared to unstructured variational methods such as mean-field VI, even when those methods are not amortized.

The rest of the paper is organized as follows. Section 2 reviews variational inference and introduce the proposed amortized structural variational inference (ASVI) framework. Section 3 provides a theoretical analysis of ASVI and applies it to a Bayesian state-space model. Section 4 presents a computational algorithm for implementing ASVI and reports results from an extensive simulation study. All proofs and additional implementation details are provided in the appendices.

2 AMORTIZED STRUCTURAL VARIATIONAL INFERENCE

Consider a dataset $X^n = (X_1, \ldots, X_n)$ consisting of n independent observations, each generated from a parametric latent variable model, with $Z^n = (Z_1, \ldots, Z_n)$ denoting a collection of local latent variables associated with individual data points. The joint likelihood of the data and latent variables, conditioned on the model parameters $\theta = (\theta_1, \ldots, \theta_d)$ shared across the dataset, is given by $p(X^n, Z^n \mid \theta) = p(Z^n \mid \theta) \cdot \prod_{i=1}^n p(X_i \mid \theta, Z_i)$, where $p(X_i \mid \theta, Z_i)$ denotes the likelihood of the individual observation X_i given the corresponding latent variable Z_i and the global parameter θ, and $p(Z^n \mid \theta)$ specifies the joint distribution of the latent variables given the model parameters. This hierarchical structure captures both shared global characteristics through θ and observation-specific variability through the local latents Z_i. Such models are common in latent variable modeling, including applications such as mixture models, topic models, and state-space models, where the latent variables encode unobserved features, cluster assignments, or temporal latent states.

In a Bayesian setting, we place a prior distribution $\pi(\theta)$ over the global parameters θ, and variational inference seeks to approximate the joint posterior distribution $p(\theta, Z^n \mid X^n)$ by selecting a distribution $\widehat{Q}$ from a chosen variational family $\mathcal{Q}$ that minimizes the Kullback–Leibler (KL) divergence to the true posterior:

$$\widehat{q} = \arg\min_{q \in \mathcal{Q}} D_{\mathrm{KL}}\big(q(\theta, Z^n) \,\big\|\, p(\theta, Z^n \mid X^n)\big), \qquad (1)$$

where $D_{\mathrm{KL}}(p \,\|\, q) = \int p \log(p/q)$ denotes the Kullback–Leibler divergence between two distributions p and q. In specifying the variational family $\mathcal{Q}$, a common choice is the mean-field family $\mathcal{Q}_{\mathrm{MF}} = \big\{q : q(\theta, Z^n) = q_\theta(\theta) \prod_{i=1}^n q_i(Z_i)\big\}$, which assumes independence between the model parameters and local latent variables, as well as across the local latents. This simplifying assumption enables efficient optimization using well-established algorithms such as coordinate ascent variational inference (CAVI, Bishop and Nasrabadi (2006)), which are computationally tractable for this factorized form. Alternatively, the Gaussian family with general covariance structure, defined as $\mathcal{Q}_{\mathrm{G}} = \big\{q : q(\theta, Z^n) = \mathcal{N}(\theta; \mu_\theta, \Sigma_\theta) \cdot \mathcal{N}(Z^n; \mu_z, \Sigma_z)\big\}$, is also widely used, with $\mathcal{N}(\,\cdot\,; \mu, \Sigma)$ denoting the (multivariate) normal distribution with mean μ and covariance matrix Σ. This family can capture correlations among parameters and can be optimized using stochastic variational inference methods Hoffman et al. (2013), which however suffers from cubic computational complexity in both the sample size and the parameter dimension due to the need to estimate and manipulate dense covariance matrices during optimization.

Amortized variational inference (AVI) (e.g., Margossian and Blei (2024)) replaces the per-sample optimization of individual variational factors $q_i(Z_i)$ with a global inference function γ that maps each observation to its corresponding variational parameters. The resulting amortized variational family takes the form:

$$\mathcal{Q}_{\mathrm{A}} = \Big\{q : q(\theta, Z^n) = q_\theta(\theta) \cdot \prod_{i=1}^n q_{\gamma_\nu(X_i)}(Z_i)\Big\}, \quad (2)$$

where γ_ν is a parameterized mapping, typically implemented using a deep neural network such as a feedforward architecture, and ν denotes its parameters.

By learning a single inference function across all data points, AVI enables efficient posterior approximation for new observations without the need to solve a separate optimization problem for each instance. A prominent application of this framework is the variational autoencoder (VAE, Tomczak and Welling (2018)), where the encoder defines the variational family $\mathcal{Q}_{\text{encoder}} = \big\{q : q(\theta, Z^n) = q_\theta(\theta) \cdot \prod_{i=1}^n \mathcal{N}(Z_i; \mu_\gamma(X_i), \sigma^2_\gamma(X_i)\big\}$, with the inference function $\gamma(x) = \big(\mu_\gamma(x),\, \sigma_\gamma(x)\big)$ modeled by deep neural networks, commonly realized as multilayer perceptrons (MLPs).

To allow each latent variable to be influenced by other observations, thereby partially capturing dependencies among latent variables, one can enrich the variational family of AVI by incorporating neighborhood information. This leads to an amortized *neighborhood-aware* variational family:

$$\mathcal{Q}_{\mathrm{AN}} = \Big\{q : q(\theta, Z^n) = q_\theta(\theta) \cdot \prod_{i=1}^n q_{\gamma(X_{C_i})}(Z_i)\Big\}, \quad (3)$$

where C_i denotes the index set of a local neighborhood associated with X_i, and X_{C_i} refers to the collection of observations indexed by C_i. In this formulation, the inference function γ depends not only on X_i but also on the neighboring observations X_{C_i}. This extension allows the variational approximation for Z_i to partially capture local structure and dependencies. Our numerical results in Section 4 show that even this simple modification can lead to a notable improvement in the performance of AVI.

It is important to note that both $\mathcal{Q}_{\mathrm{A}}$ and $\mathcal{Q}_{\mathrm{AN}}$ remain within the mean-field family, as they assume conditional independence among latent variables. Although neighborhood information is used in the inference function γ, these families do not encode dependencies directly in the variational distribution. This limitation becomes significant in models with inherent latent structure (such as hidden Markov models, state-space models, or latent graphical models), where ignoring such structure can lead to inconsistent estimation Wang et al. (2022). This motivates the development of the *amortized structural variational inference* (ASVI) framework, which incorporates structural dependencies into both the variational family and the inference function.

Specifically, We define a dependency-aware version of the variational family, denoted $\mathcal{Q}_{\mathrm{AS}}$, that explicitly integrates structural information into the joint approximation over latent variables:

$$\mathcal{Q}_{\mathrm{AS}} = \Big\{q : q(\theta, Z^n) = q_\theta(\theta) \cdot q_\gamma(Z^n)\Big\}, \quad (4)$$

where $q_\gamma(Z^n)$ preserves the dependency structure present in the joint latent variable distribution $p(Z^n \,|\, \theta)$ Wang et al. (2022), and γ denotes the corresponding inference function, which also incorporates neighborhood information as described in (3). For example, suppose the latent distribution factorizes over cliques clique(G) in a graphical model, $p(Z^n \,|\, \theta) = \prod_C p_C(Z_C \,|\, \theta)$, where C ranges over all cliques in clique(G), and $Z_C := (Z_i : i \in C)$ denotes the subset of latent variables indexed by C. In this case, one can define an amortized variational family that factorizes over the same cliques: $q_\gamma(Z^n) = \prod_C q_{\gamma(X_C)}(Z_C)$, where γ maps the observations within each clique X_C to the variational parameters for the corresponding latent variables Z_C.

For concreteness, in this paper, we focus on latent variable models where the *latent structure is governed by a graph*. Specifically, we consider the latent variable model:

$$X_i \mid Z_i \sim p_\mu(X_i \mid Z_i), \;\; i = 1, \ldots, n,$$
$$Z^n = (Z_1, \ldots, Z_n) \sim p_\lambda(Z_1, \ldots, Z_n),$$

where $\theta = (\mu, \lambda)$ are the global model parameters. The latent distribution p_λ is structured according to a graph $G = (V, E)$, with $\mathcal{N} = \{C_1, \ldots, C_n\}$ denoting the neighborhood system induced by G, so that Z satisfies a local Markov property with respect to G; that is, $Z_i \perp Z_j \mid$

$Z_{C_i \setminus \{i\}}$ for all $j \notin C_i$. In this setting, ASVI can be used to amortize the local parameters of the conditional distributions $q(Z_i \mid Z_{C_i \setminus \{i\}})$ in the variational family $\mathcal{Q}_{\text{AS}}$, using the corresponding observations x_{C_i}.

To illustrate $\mathcal{Q}_{\text{AS}}$, we focus on general state-space models Koller and Friedman (2009); Zeng (2013), specified by

$$X_t \mid Z_t \sim p_\mu(X_t \mid Z_t) \text{ and } Z_t \mid Z_{t-1} \sim p_\lambda(Z_t \mid Z_{t-1}),$$

where the latent variables follow a first-order Markov structure. In this case, the neighborhood for each observation is defined as $C_t = \{t-1, t, t+1\}$, except at the boundaries (e.g., $C_1 = \{1, 2\}$).

We begin with a structured variational family that reflects the Markov structure of the latent state-space model. Before introducing the *amortized structural variational family* $\mathcal{Q}_{\text{AS}}$, we first define a structured variational distribution as

$$q(\theta, Z^n) = q_\theta(\theta) \cdot q_{\phi_1}(Z_1) \cdot q_{\phi_2}(Z_2 \mid Z_1) \cdots q_{\phi_n}(Z_n \mid Z_{n-1}).$$

ASVI amortizes the variational parameters $\phi^n = (\phi_1, \ldots, \phi_n)$ of the latent variables by mapping the local neighborhoods X_{C_i} to the parameters of the corresponding conditional distributions. The resulting *amortized structural variational family* is defined as

$$\mathcal{Q}_{\text{AS}} = \Big\{ q : q(\theta, Z^n) = q_\theta(\theta) \cdot q_{\gamma(X_{C_1})}(Z_1) \cdot q_{\gamma(X_{C_2})}(Z_2 \mid Z_1) \cdots q_{\gamma(X_{C_n})}(Z_n \mid Z_{n-1}) \Big\}.$$

Here, the inference function γ is designed to incorporate contextual information from neighboring observations and may be modeled using deep neural networks, such as MLPs.

3 THEORETICAL RESULTS

In this section, we study the theoretical properties of the proposed ASVI framework through the lens of variational risk bounds, and then apply these results to latent state-space models. Our goal is to evaluate the quality of the amortized variational posterior $\widehat{q} = \widehat{q}_\theta \times \widehat{q}_{Z^n}$ obtained by minimizing the VI objective (1) using the structured inference family $\mathcal{Q}_{\text{AS}}$ (see equation (4)). To this end, we derive a nonasymptotic upper bound on the amortized variational risk. Working under a frequentist setting in which the data X^n is generated from our considered parametric latent variable model with true parameter θ^*, we show that the estimated marginal posterior $\widehat{q}_\theta$ concentrates around θ^* under appropriate discrepancy measures as sample size n grows. To simplify the theoretical analysis, we follow Yang et al. (2020); Wang et al. (2022) and adopt the α-variational inference framework (See Appendix B for further details), which reduces the analysis by requiring verification of a minimal number of conditions.

In this framework, the risk function is defined using the α-Rényi divergence D_α Van Erven and Harremos (2014), where $D_\alpha(q, p) = \frac{1}{\alpha - 1} \log \int q^\alpha p^{1-\alpha}$ for two distribution p and q. When $\alpha = 0.5$, the Rényi divergence corresponds to the squared Hellinger distance, while letting $\alpha \to 1_-$ recovers the KL divergence, which is commonly used in standard VI. More specifically, we use a sample size rescaled α-Rényi divergence $D_\alpha^{(n)}(\theta, \theta^*) = n^{-1} D_\alpha\big(p_\theta^{(n)}, p_{\theta^*}^{(n)}\big)$ as a measure of discrepancy between θ and θ^*, where $p_\theta^{(n)}$ denotes the marginal density of X^n under parameter θ by integrating out the latent variables, that is, $p_\theta^{(n)}(X^n) = \int p(X^n \mid Z^n, \theta)\, p(Z^n \mid \theta)\, dZ^n$. See Section 3 of Wang et al. (2022) for further discussions.

We begin by establishing a general bound for an arbitrary amortized variational family and then specialize the result to the state-space model setting with concrete emission and transition structures.

Throughout, we let $\mathcal{M}$ and Λ denote the parameter spaces for the observation distribution p_μ and the latent variable distribution p_λ, respectively.

To begin with, we have the following definitions. Let $\widehat{q}_{\theta,\alpha}$ denote the amortized variational posterior obtained from the α-variational inference framework (see Appendix B) using the variational family $\mathcal{Q}_{\text{AS}}$, consisting of distributions of the form $q(Z^n, \theta) = q_{\gamma(X^n)}(Z^n) \cdot q_\theta(\theta)$, where γ is a deterministic inference function over X^n.

Definition 1 *Let $\pi^*_{Z^n} := p(Z^n \mid X^n, \theta^*)$ denote the latent posterior distribution under true parameter $\theta^* = (\lambda^*, \mu^*)$, and $\Delta_{ASVIGap}$ denote the amortization gap defined as*

$$\Delta^2_{ASVIGap} = \inf_{q_{\gamma(X^n)} \in \mathcal{Q}_{AS}} n^{-1} D\left(q_{\gamma(X^n)} \,\|\, \pi^*_{Z^n}\right). \quad (5)$$

Definition 2 *The variational gap Δ_{VIGap} is defined by*

$$\Delta^2_{VIGap} = D\alpha \left[\frac{f_\lambda(n)}{n} \varepsilon_\lambda^2 + f_\mu(n) \varepsilon_\mu^2 \right] - \frac{1}{n} \Big(\log P_\lambda \left(\mathcal{B}_n^{\text{VI}}(\lambda^*, \varepsilon_\lambda) \right) + \log P_\mu \left(\mathcal{B}_n^{\text{VI}}(\mu^*, \varepsilon_\mu) \right) \Big).$$

Here, functions f_μ and f_λ are defined in the neighborhoods $\mathcal{B}_n^{\text{VI}}(\lambda^*, \varepsilon_\lambda)$ and $\mathcal{B}_n^{\text{VI}}(\mu^*, \varepsilon_\mu)$ through

$$\left\{ \lambda \in \Lambda : \begin{array}{l} D\left(p(Z^n \mid \lambda^*) \,\|\, p(Z^n \mid \lambda)\right) \le f_\lambda(n) \varepsilon_\lambda^2 \\ V\left(p(Z^n \mid \lambda^*) \,\|\, p(Z^n \mid \lambda)\right) \le f_\lambda(n) \varepsilon_\lambda^2 \end{array} \right\}, \quad (6)$$

$$\left\{ \mu \in \mathcal{M} : \begin{array}{l} \max\limits_{1 \le i \le n} \mathbb{E}_{Z^n \mid \theta^*} D_i(\mu^*, \mu) \le f_\mu(n) \varepsilon_\mu^2 \\ \max\limits_{1 \le i \le n} \mathbb{E}_{Z^n \mid \theta^*} V_i(\mu^*, \mu) \le f_\mu(n) \varepsilon_\mu^2 \end{array} \right\}, \quad (7)$$

with $D_i(\mu^*, \mu) = D\left[p(\cdot \mid \mu^*, X_i) \,\|\, p(\cdot \mid \mu, X_i)\right]$ and $V_i(\mu^*, \mu) = V\left[p(\cdot \mid \mu^*, X_i) \,\|\, p(\cdot \mid \mu, X_i)\right]$, where we

used the shorthand of $D(p\,\|\,q) := \int p\log(p/q)d\mu$ *and* $V(p\,\|\,q) = \int p\log^2(p/q)d\mu$ *for two distributions* p *and* q.

Theorem 1 *Then, for any fixed* $(\varepsilon_\lambda, \varepsilon_\mu) \in (0,1)^2$ *and* $D > 1$, *with probability at least* $1 - \frac{5}{(D-1)^2\left(f_\lambda(n)\varepsilon_\lambda^2 + n f_\mu(n)\varepsilon_\mu^2\right)}$, *it holds that*

$$\int D_\alpha^{(n)}(\theta,\theta^*)\,\widehat{q}_{\theta,\alpha}(d\theta) \le \frac{1}{1-\alpha}\left(D\,\alpha\,\Delta^2_{\mathit{ASVIGap}} + \Delta^2_{\mathit{VIGap}}\right).$$

Theorem 1 provides a non-asymptotic upper bound on the variational risk incurred by amortized inference using a structured inference function γ. The bound decomposes into three interpretable components: (i) the divergence and variance terms $f_\lambda(n)\varepsilon_\lambda^2$ and $f_\mu(n)\varepsilon_\mu^2$, which correspond to the standard variational approximation error; (ii) two logarithmic penalty terms that control concentration around the true parameters λ^* and μ^*; and (iii) the amortization gap term $\Delta^2_{\text{ASVIGap}}$, which quantifies the additional error introduced by amortization. In particular, the first two components define the variational gap, which characterizes the approximation error arising from the use of the structured variational family $\mathcal{Q}_S = \left\{q : q(\theta, Z^n) = q_\theta(\theta)\cdot q_{Z^n}(Z^n)\right\}$.

The variational risk bound in Theorem 1 explicitly highlights how the design and expressiveness of the inference function γ influence the overall approximation quality. In particular, our result generalizes Theorem 1 of Wang et al. (2022), which analyzes the variational risk for non-amortized posteriors based on structured variational families. The additional term $\Delta^2_{\text{ASVIGap}}$ in equation (5) quantifies the amortization gap, reflecting the cost of using a shared inference function γ instead of optimizing each variational factor independently. Specifically, $\Delta^2_{\text{ASVIGap}}$ measures the approximation error of approximating the latent variables posterior $p(Z^n \mid X^n, \theta^*)$ under θ^* using the amortized variational family $\mathcal{Q}_{\text{AS}}$. When this gap is negligible, such as in cases with high inference function capacity or favorable model structure, our bound reduces to the non-amortized result as a special case.

We now apply Theorem 1 in the context of a multivariate latent state-space model, where $Z_i \in \mathbb{R}^d$.

This model exhibits a first-order Markov structure in the latent sequence $(Z_1, \ldots, Z_n)$, which is naturally leveraged by the ASVI framework. We adopt the structured amortized variational family $\mathcal{Q}_{\text{AS}}$, which respects the latent Markovian dependence through the factorization

$$q(\theta, Z^n) = q(\theta)\, q_{\gamma(X_1)}(Z_1) \prod_{i=2}^{n} q_{\gamma(X_i)}(Z_i \mid Z_{i-1}),$$

where each conditional factor is truncated to maintain bounded support:

$$q_{\gamma(X_i)}(Z_i \mid Z_{i-1}) = \mathcal{TN}_{[-R_1,R_1]^d}\left(Z_i;\, \mu,\, \Sigma\right),$$

where $\mu = A_\gamma(X_i)\, Z_{i-1} + b_\gamma(X_i)$ and $\Sigma = \Sigma_\gamma(X_i)\left(I_d + 2\cdot\frac{Z_{i-1}Z_{i-1}^\top}{1+\|Z_{i-1}\|_2^2}\right)$.
The inference function $\gamma : \mathbb{R}^d \to \mathbb{R}^{d\times d} \times \mathbb{R}^d \times \mathbb{S}^d_{++}$ maps each observation X_i to a triplet $(A_\gamma(X_i), b_\gamma(X_i), \Sigma_\gamma(X_i))$, where $\mathbb{S}^d_{++}$ denotes the space of all d-by-d positive definite matrices. Under mild regularity conditions on the outputs of γ, we obtain the following high-probability bound on the variational risk for the resulting amortized variational posterior $\widehat{q}_{\theta,\alpha}$.

Corollary 1 *Consider the truncated state space model described above. Suppose the inference function is implemented as a fully connected ReLU neural network of depth* L *and width* r, *this is* $\gamma \in \mathcal{F}(L, r)$, *see Appendix A. We assume either of the following neural network configurations,* $L \asymp \log(n)$, $r \asymp (n)^{1/[2(2p+1)]}$, *or* $L \asymp \log(n)\cdot(n)^{1/[2(2p+1)]}$, $r = O(1)$, *for a sufficiently large* $p > 0$. *Let* $m := d^2 + 2d$ *denote the total number of scalar-valued inference function outputs, corresponding to the entries of* $A_\gamma(x), b_\gamma(x)$, *and the diagonal of* $\Sigma_\gamma(x)$. *Assume the setup and notation from Theorem 1, and define* $\varepsilon_\lambda = \varepsilon_\mu = \frac{(\log n)^\beta}{n}$ *for some* $\beta > 0$. *Then it follows that* $f_\lambda(n) = O(n)$, $f_\mu(n) = O(1)$, *and the amortization gap satisfies* $\Delta^2_{\mathit{ASVIGap}} \lesssim m\cdot\log^c(n)\, n^{-\frac{2p}{2p+1}}$. *As a consequence, there exist constants* $C(R_1,R_2) > 0$ *and* $\beta' > 0$ *such that, with probability approaching to one under the true parameter* θ^*, *the amortized variational posterior satisfies:*

$$\int D_\alpha^{(n)}(\theta,\theta^*)\,\widehat{q}_{\theta,\alpha}(d\theta) \le$$
$$C(R_1,R_2)\, D\left(\frac{(\log n)^{\beta'}}{n} + m\cdot\log^c(n)\, n^{-\frac{2p}{2p+1}}\right),$$

where $c > 0$ *is a universal constant.*

Corollary 1 illustrates the application of general results in Theorem 1 to amortized inference in linear state-space models with truncated Gaussian innovations. In this setting, we use a structured variational family $\mathcal{Q}_{\text{AS}}$ that captures temporal dependencies through conditionals of the form $q(Z_i \mid Z_{i-1})$, where the variational parameters are generated by neural inference functions $\gamma(X_i)$. Here, we use a structured variational family that preserves the autoregressive structure, as it is known that the mean-field family, which completely ignores dependencies among latent variables, can lead to inconsistent estimation of θ Wang et al. (2022). Moreover, Corollary 1 suggests that, for this example, setting the neighborhood size to one (i.e., including only X_i) is sufficient to control the amortization gap Δ_{ASVIGap}. Our numerical results in Section 4 demonstrate that including additional neighbors does not lead to noticeable improvement in the variational approximation, which is consistent with our theoretical prediction.

Compared to Wang et al. (2022), which analyzes linear Gaussian state-space models using non-amortized, fully

factorized variational approximations, our framework applies to substantially more general settings. We allow for non-linear models, amortized inference via context-aware inference networks, and structured variational families that respect latent Markovian dependencies. While Wang et al. (2022) establishes a variational risk bound of order $\mathcal{O}(1/n)$ under strong assumptions, our bound retains the same leading $\mathcal{O}(1/n)$ term, augmented by an explicit amortization error of order $\mathcal{O}(n^{-2p/(2p+1)} \log^c n)$. This additional term captures the approximation quality of the inference function and vanishes at a near-parametric rate under mild smoothness conditions on the true parameter functions.

Overall, our result decomposes the variational risk into two components: (i) statistical estimation error due to the structured variational approximation and (ii) amortization error arising from learning the inference function. This decomposition highlights a fundamental tradeoff in amortized variational inference: improved scalability and generalization are achieved at the cost of an additional approximation gap. However, this gap can be explicitly quantified under regularity assumptions and diminishes rapidly as smoothness increases. In particular, as $p \to \infty$, the amortization gap approaches $\mathcal{O}(1/n)$, thereby recovering the fully optimized variational rate. By explicitly bounding the variational risk in terms of smoothness assumptions and network complexity, Corollary 1 highlights the practical relevance of Theorem 1. It demonstrates how structural modeling choices and neural network architecture together govern the statistical efficiency of modern amortized variational inference methods.

4 SIMULATION STUDY AND REAL DATA ANALYSIS

In this section, we present our algorithms for Amortized Neighbor Variational Inference (ANVI) and Amortized Structured Variational Inference (ASVI), and evaluate their performance through a numerical study. Code repository is available at: https://github.com/waterism211/Amortized-Structured-Variational-Inference. We examine how incorporating structural information (either through a structured variational family or a structure-aware inference mapping) improve VI performance. To this end, we generate data from latent variable models with either a Markov structure or a latent graph structure, and apply our algorithms accordingly. Our results demonstrate improvements in both computational efficiency and estimation accuracy, as measured by reduction in run-time and increases in the Evidence Lower Bound (ELBO), correspondingly, compared to several state-of-the-art methods.

Table 1: Variational families considered in this work: MF = mean-field VI; Const = naive amortized VI; A = amortized VI; AN = neighborhood-based amortized VI; S = structured VI; AS = amortized structured VI.

Family	$q(\theta, Z^n)$ factorization
$\mathcal{Q}_{MF}$	$q_0(\theta) \prod_{i=1}^n q_i(Z_i)$
$\mathcal{Q}_{Const}$	$q_0(\theta) \prod_{i=1}^n q_{\text{const}}(Z_i)$
$\mathcal{Q}_{A}$	$q_0(\theta) \prod_{i=1}^n q_{\gamma(X_i)}(Z_i)$
$\mathcal{Q}_{AN}$	$q_0(\theta) \prod_{i=1}^n q_{\gamma(X_{C_i})}(Z_i)$
$\mathcal{Q}_{S}$	$q_0(\theta)\, q_1(Z_1) \prod_{i=2}^n q_i(Z_i \mid Z_{i-1})$
$\mathcal{Q}_{AS}$	$q_0(\theta)\, q_{\gamma(X_{C_1})}(Z_1) \prod_{i=2}^n q_{\gamma(X_{C_i})}(Z_i \mid Z_{i-1})$

4.1 Variational Families

We consider the variational families summarized in Table 1 to approximate the joint posterior of (θ, Z^n).

Here C_i denotes a neighborhood of X_i with $|C_i| > 1$. The families satisfy $\mathcal{Q}_{\text{const}} \subset \mathcal{Q}_{\text{A}} \subset \mathcal{Q}_{\text{AN}} \subset \mathcal{Q}_{\text{MF}}$ and $\mathcal{Q}_{\text{AS}} \subset \mathcal{Q}_{\text{S}}$. All variational distributions are Gaussian, and the amortization map $\gamma(\cdot)$ is modeled by a two-layer ReLU network. For $\mathcal{Q}_{\text{A}}$ and $\mathcal{Q}_{\text{AN}}$, we use $q_{\gamma(\cdot)}(z) = \mathcal{N}(z; \mu_\gamma(\cdot), \Sigma_\gamma(\cdot))$. For $\mathcal{Q}_{\text{S}}$, the latent variables follow a linear Gaussian state-space model $q(Z^n) = \mathcal{N}(Z_1; b_1, \Sigma_1) \prod_{i=2}^n \mathcal{N}(Z_i; b_i + A_{i-1} Z_{i-1}, \Sigma_i)$. For $\mathcal{Q}_{\text{AS}}$, the parameters are amortized, yielding $q(Z^n) = \mathcal{N}(Z_1; A_\gamma(X_{C_1}), \Sigma_\gamma(X_{C_1})) \prod_{i=2}^n \mathcal{N}(Z_i; b_\gamma(X_{C_i}) + A_\gamma(X_{C_i}) Z_{i-1}, \Sigma_\gamma(X_{C_i}))$. All methods use a fractional posterior with $\alpha = 0.99$.

4.2 Algorithms for ANVI and ASVI

The amortization map is learned by maximizing the ELBO

$$\mathcal{L} = \mathbb{E}_{q(\theta, Z^n)}[\log p_\alpha(X^n, Z^n, \theta) - \log q(\theta, Z^n \mid X^n)],$$

where $p_\alpha(X^n, Z^n, \theta) = p^\alpha(X^n \mid Z^n, \theta)\, \pi(Z^n \mid \theta)\, \pi(\theta)$ and $q(\theta, Z^n \mid X^n) = q(\theta \mid X^n)\, q(Z^n \mid \theta, X^n)$. Algorithm 1 summarizes the ASVI procedure, while the ANVI algorithm is deferred to Appendix C. The only difference lies in the Monte Carlo sampling step: in ASVI the latent variables Z^n are sampled conditionally, whereas in ANVI they are sampled jointly.

4.3 AR(1) model

As our first example, we consider the simple AR(1) latent variable model and conduct a comprehensive simulation study to demonstrate that the ASVI framework is computationally efficient. Our results indicate that incorporating structural information improves estimation accuracy by reducing the amortization gap. Moreover, adopting the ASVI variational family provides a more expressive approximation, which reduces the variational gap from the outset.

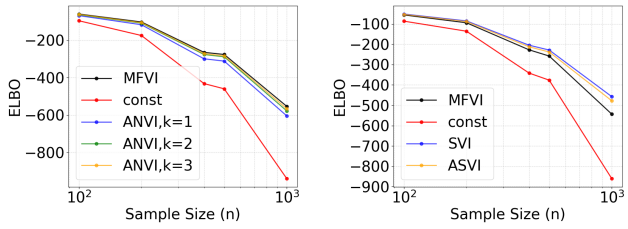
Algorithm 1 Amortized Structured Variational Inference

Input: Observations $X^n = \{X_1, \ldots, X_n\}$.
Output: Variational parameters $\{A_z, b_z, \Sigma_z, \mu_\theta, \Sigma_\theta\}$ defining $q(Z_1) = \mathcal{N}(b_z, \Sigma_z)$, $q(Z_i \mid Z_{i-1}) = \mathcal{N}(b_z + A_z Z_{i-1}, \Sigma_z)$ for $i \geq 2$, and $q(\theta) = \mathcal{N}(\mu_\theta, \Sigma_\theta)$.
Initialize $\{A_z, b_z, \Sigma_z, \mu_\theta, \Sigma_\theta\}$.
while the loss $\mathcal{L}(X^n; \theta, Z^n)$ has not converged **do**
    Sample $\{\theta_j\}_{j=1}^m \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$.
    for $i = 1, \ldots, n$ **do**
        Sample $Z_{1j} \sim q(Z_1)$ and $Z_{ij} \sim q(Z_i \mid Z_{i-1,j})$ for $j = 1, \ldots, m$.
        Compute $\mathcal{L}(X_i; \theta, Z_i) = \sum_{j=1}^m [\log p_\alpha(X_i, Z_{ij}, \theta_j) - \log q(\theta_j, Z_{ij} \mid X_i)]$.
        Compute gradients $\nabla \mathcal{L}(X_i; \theta, Z_i)$ w.r.t. $\{A_z, b_z, \Sigma_z, \mu_\theta, \Sigma_\theta\}$.
    end for
    $\widehat{\nabla}\mathcal{L} = \frac{1}{n} \sum_{i=1}^n \nabla \mathcal{L}(X_i; \theta, Z_i)$.
    Update parameters using $\widehat{\nabla}\mathcal{L}$.
end while
return $\{A_z, b_z, \Sigma_z, \mu_\theta, \Sigma_\theta\}$.

We begin by simulating data from a hidden Markov model with initial state $Z_1 \sim \mathcal{N}(0, 1)$. For $i = 2, \ldots, n$, the data-generating process is defined as:

$$Z_i = 0.5 Z_{i-1} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \tau^2),$$
$$X_i = \theta + \sin(Z_i) + \eta_i, \quad \eta_i \sim \mathcal{N}(0, \sigma^2).$$

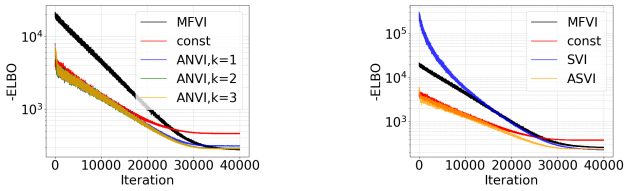
Here we set the global parameter to be learned as $\theta = 2$, with fixed noise levels $\tau = 0.5$ and $\sigma = 0.7$, and simulate datasets of varying sample sizes under this model specification. Figure 1 shows the ELBO values for different methods as a function of sample size.



(a) ELBO of the ANVI methods (b) ELBO of the ASVI methods

Figure 1: The ELBO for different VI methods versus sample size n. Here k stands for the number of neighbors used in the inference function γ. A larger ELBO value indicates better performance.

In Figure 1a, we focus on the mean-field variational family, with the goal of illustrating that incorporating structural information into the inference map improves accuracy, as reflected in larger ELBO values. The ELBO values for the light orange line (ANVI with $k = 3$, using three neighbors) and the green line (ANVI with $k = 2$, using two neighbors) are higher than those of the blue line, which corresponds to the amortized algorithm without neighborhood information (ANVI with $k = 1$). Even the blue line (ANVI with $k = 1$) demonstrates a clear improvement over the constant variational family. The performance difference between ANVI with $k = 3$ and $k = 2$ is smaller, although ANVI with $k = 3$ still performs slightly better. This finding aligns with the fact that the data were generated from a process involving two neighbors. In Figure 1b, we examine structured



(a) ANVI (b) ASVI

Figure 2: The optimization sample paths of different methods. A small value at the final iteration is preferred.

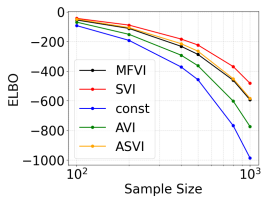
variational families. The ELBO achieved by ASVI (light orange line) is higher than that of mean-field variational inference (MFVI) without amortization (black line). This demonstrates that the ASVI family not only reduces the amortization error but also yields a tighter variational approximation, leading to a smaller variational error compared to the standard mean-field class.

In assessing computational efficiency, Figure 2a shows that ASVI mainly improves the early stage of optimization because it requires fewer burn-in steps, rather than reducing the total number of epochs needed for convergence. Figure 2b presents a similar pattern.

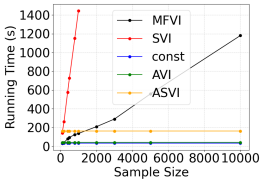
In Figure 1, the algorithm is retrained for each sample size using different datasets corresponding to varying values of n. This setup illustrates that access to more data leads to larger reductions in the amortization gap. By contrast, Figure 3a fixes a training set with $n = 100$ observations to learn the inference map, which is then applied to evaluate ELBO values as new data arrive—without retraining from scratch. Despite this difference, Figure 3a shows a pattern consistent with Figure 1: ASVI (light orange line) consistently achieves the highest ELBO among the amortized methods and even outperforms MFVI without amortization (black line).

Figure 3b reports running time as a function of sample size n. The running time of ASVI (light orange line) remains essentially *constant* as n grows, in sharp contrast to the linear increases observed for MFVI (black) and SVI (red). This demonstrates the superior scalability of the amortized inference provided by ASVI.

We conduct an additional experiment in which ASVI is fully trained for each sample size n. As shown in Table 2, the training time of ASVI grows approximately linearly with the



(a) ELBO as a function of sample size



(b) Running time as a function of sample size

Figure 3: ELBO and running time as a function of n

training sample size n_{train}. This scaling behavior is consistent with the theoretical linear complexity of our update rule and confirms that ASVI remains computationally tractable even when the amortizer is retrained for each n. The predictive MSE and ELBO values are also stable across different training and test sizes, indicating that performance does not degrade with larger n. Overall, these results suggest that ASVI remains competitive with MFVI and SVI in terms of computational cost when training is performed separately for each sample size.

Our theory suggests that the overall approximation error decomposes into two components: the amortization gap and the variational gap. When the amortizer is pre-trained at a fixed sample size n_{train}, the amortization gap remains approximately constant as more data arrive, while the variational gap decreases with n_{test}, since parameter estimation under the variational approximation improves with larger datasets. This implies a characteristic pattern: as n_{test} increases, the overall error initially decreases (when the variational gap dominates) and eventually stabilizes once the amortization gap becomes dominant.

The empirical results in Table 2 are consistent with this behavior. Here the MSE(pred) $= \frac{1}{n}\sum_{i=1}^{n}\|X_i - \widehat{X}_i\|_2^2$, where $\widehat{X}_i$ is the prediction by the posterior mean of Z_i. For example, when the amortizer is trained at $n_{\text{train}} = 300$, the predictive error MSE(pred) decreases as n_{test} increases from 100 to 400, reflecting the shrinking variational gap, and then stabilizes around $n_{\text{test}} = 500$, where the amortization gap dominates. Moreover, across rows, increasing the pre-training size consistently improves performance: larger n_{train} yields uniformly better ELBO and MSE(pred) values, as expected from the reduced amortization gap. Empirically, we do not observe significant degradation in performance as more data arrive; instead, performance improves until it reaches the limit imposed by the amortization gap.

4.4 AR(p) model

In this example, we allow for model misspecification by generating data from a true process with a long dependence window, as the following AR(64) process: However, the statistical model we fit assumes an AR(3) structure for the latent variables $\{Z_i\}_{i=1}^n$, where both the latent process and the observed data depend on the model parameters. In other words, when fitting the model we restrict the latent dynamics to an AR(3) process by setting $c_k \equiv 0$.

$$Z_t = \sum_{j=1}^{3} a_j Z_{t-j} + \sum_{k=4}^{K_{\max}} c_k Z_{t-k} + \varepsilon_t,\ \varepsilon_t \sim \mathcal{N}(0, \tau^2 I_3),$$

$$X_t = \theta + Z_t + \sum_{j=1}^{3} b_j X_{t-j} + \eta_t,\ \eta_t \sim \mathcal{N}(0, \sigma^2 I_3).$$

In this simulation study, we compare ASVI with different amortization window sizes k for the variational family $\mathcal{Q}_{AS} = \{q : q(\theta, Z^n) = q_0(\theta)\, q_1(Z_1, Z_2, Z_3) \cdot \prod_{i=4}^{n} q_{\gamma(X_{i:i+k})}(Z_i \mid Z_{i-1}, Z_{i-2}, Z_{i-3})\}$. For benchmarking, we also applied flow-based methods (Rezende and Mohamed, 2015). We will use two more metrics to compare our methods. The MSE(global) is the mean square error of posterior means of global parameters, and the MSE(pred) $= \frac{1}{n}\sum_{i=1}^{n}\|X_i - \widehat{X}_i\|_2^2$, where $\widehat{X}_i$ is the prediction by the posterior mean of Z_i. MSE(pred) is computed using causal predictions, meaning that the prediction at time n is based solely on observations up to time $n-1$. Results from Table 3 show that MSE(global) does not decrease monotonically as the amortization window k increases. In fact, the smallest MSE(global) occurs at $k = 2$, which corresponds to the minimal window size needed to capture the posterior dependence. The increase in MSE(global) for larger k may be due to overfitting, since a larger neighborhood requires estimating more parameters to model local dependence, which can reduce the efficiency of parameter estimation, as reflected in MSE(global). In comparison, MSE(pred) is the predictive error, which is largely determined by the noise level σ (which is fixed) in the data-generating model and varies only slightly across ASVI variants.

Table 2: ASVI ELBO, MSE(pred), and training time across training and test sizes.

	Test size n_{test}					
Metric / n_{train}	100	200	300	400	500	Time (s)
ELBO, 100	-135.60	-278.25	-432.96	-548.79	-712.59	89.63
ELBO, 200	-139.53	-281.06	-407.08	-543.52	-682.04	163.69
ELBO, 300	-143.02	-282.86	-403.33	-531.19	-670.01	243.65
ELBO, 400	-140.68	-279.78	-403.30	-522.61	-672.32	323.68
ELBO, 500	-143.36	-277.91	-398.10	-531.43	-676.93	410.66
MSE(pred), 100	0.3956	0.4225	0.4316	0.4122	0.4277	
MSE(pred), 200	0.3956	0.4083	0.3994	0.3832	0.3956	
MSE(pred), 300	0.4083	0.4058	0.3733	0.3588	0.3709	
MSE(pred), 400	0.3844	0.3906	0.3636	0.3481	0.3588	
MSE(pred), 500	0.3856	0.3869	0.3600	0.3446	0.3540	

4.5 Nonlinear AR(2) model

In this example, we consider the following challenging nonlinear latent process where $Z_i \in \mathbb{R}^2$ and $X_i \in \mathbb{R}^2$, with a

Table 3: AR(p) performance summary. Global parameters are $a_1, a_2, a_3, b_1, b_2, b_3, \theta$

Method	ELBO	MSE (global)	MSE (pred)	Time (s)
MFVI	-873.48	0.2254	0.3272	28.5
SVI	**-811.25**	**0.0082**	**0.3147**	1282.3
ASVI_k1	-833.40	0.0298	0.3564	1241.9
ASVI_k2	-835.72	0.0267	0.3387	1262.7
ASVI_k3	-816.71	0.0314	0.3249	1240.8
FlowASVI_k1	-838.07	0.0948	0.3624	1278.6
FlowASVI_k2	-835.95	0.0996	0.3564	1277.9
FlowASVI_k3	-835.40	0.1048	0.3376	1258.9

nonlinear dynamic relationship among the latent variables. This setting is substantially more complex than the linear case, where standard Kalman filter–type algorithms are no longer applicable. So we still persist the SVI as $q_i(Z_i \mid Z_{i-1}, Z_{i-2}) = \mathcal{N}(Z_i; b_i + A_{i-1}^{(1)} Z_{i-1} + A_{i-2}^{(2)} Z_{i-2}, \Sigma_i)$.

$$p(X_i \mid Z_i;\, c, \tau^2) = \mathcal{N}\big(X_i;\; c\, Z_i,\; \tau^2 I\big),$$
$$p(Z_i \mid Z_{i-1}, Z_{i-2};\, \theta) = \mathcal{N}\big(Z_i;\; \mu_i,\; \sigma^2 I\big),$$
$$\mu_i = a_1 \odot \tanh\big(b_1 \odot Z_{i-1}\big) + a_2 \odot \tanh\big(b_2 \odot Z_{i-2}\big).$$

To capture nonlinear dependencies in the latent dynamics within ASVI, we augment the inference function γ with an additional MLP that maps (Z_{i-1}, Z_{i-2}, X_i) to the local variational parameters. Specifically, the mean of $q_{\gamma(X_i)}(Z_i \mid Z_{i-1}, Z_{i-2})$ is parameterized as $\mathrm{MLP}_\mu\big([Z_{i-1}, Z_{i-2}, X_i]\big)$. This amortized form enables q to flexibly model nonlinear interactions between Z_{i-1} and Z_{i-2} while conditioning on the local observation X_i.

Table 4: Nonlinear AR(2) performance summary. Global parameters are a_1, a_2, b_1, b_2, c

Method	ELBO	MSE (global)	MSE (pred)	Time (s)
MFVI	-1153.21	1.6869	0.0146	904.63
SVI	-215.95	0.1861	0.0047	1171.17
ASVI	-205.02	**0.0255**	**0.0027**	1438.21
FlowASVI	**-201.31**	0.0815	0.0249	2761.24

Across Table 4, ASVI consistently outperforms SVI: with its flexible amortized MLP, ASVI achieves higher ELBO, lower latent MSE(pred), and smaller MSE(global). In addition, ASVI surpasses Flow-ASVI in both prediction accuracy and parameter estimation.

4.6 Real data example

We consider Moving-MNIST sequences $X_{1:N}$, where each frame $X_i \in [0,1]^{64\times 64}$ is a grayscale image. We learn latent representations $Z_i \in \mathbb{R}^d$ with $d = 64$ using an encoder–decoder architecture. For evaluation, we compare one-step priors over VAE latents on Moving-MNIST based on a pretrained convolutional VAE. For prediction, we train the conditional prior $q(Z_i|Z_{i-1})$, and decode it using the pretrained VAE. Our ASVI-based prior is implemented as a MLP that takes the most recent latents and predicts the next one. During training, we introduce a guide network that has access to the current image X_i along with the latent window, and distills this information into the prior. At test time, however, the prior operates solely on past latents without access to X_i.

We compare our approach with three baselines: (i) *Random*, where the next latent is drawn from a standard normal; (ii) *Gaussian*, where an MLP predicts the next latent under a Gaussian prior; and (iii) *Flow*, where a RealNVP prior is conditioned on the latent window. Experiments use the standard Moving-MNIST test split. For each sequence, we evaluate multiple time steps and report metrics averaged across all pairs. Performance is compared based on predictive mean-square error (MSE). As shown in Table 5, ASVI achieves the lowest test MSE among all priors, demonstrating superior one-step forecasting.

Prior	Train MSE	Test MSE
Random	0.0540	0.0545
ASVI	0.0150	**0.0176**
Gaussian	0.0181	0.0198
Flow	**0.0060**	0.0206

Table 5: One-step prediction MSE comparison.

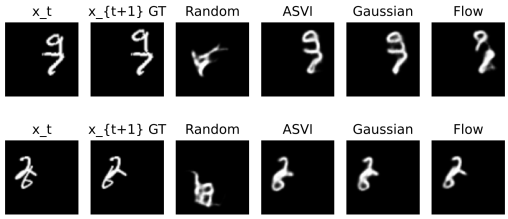


Figure 4: Bottom row: example of predicted moving digit from its previous frame using different methods on the training dataset. Top row: corresponding results on the test dataset.

5 CONCLUSION & DISCUSSION

In this work, we propose a novel amortized structured variational inference (ASVI) framework that employs a structured variational family parameterized by an inference (amortization) map. The amortization map is implemented using neural architectures that encode additional local neighborhood structure, beyond that induced by the model. ASVI is scalable and comes with theoretical guarantees, reducing both the variational gap and the amortization gap. Future work will focus on extending ASVI to more complex modeling settings, for example, observations with dependence structures beyond conditional independence given Z and θ. Other directions include developing ASVI frameworks that can learn the underlying local neighborhood dependence structure of the data, as well as extending amortized inference to full posterior approximation.

References

Agrawal, A. and Domke, J. (2021). Amortized variational inference for simple hierarchical models. *Advances in Neural Information Processing Systems*, 34:21388–21399.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*. Springer.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Cremer, C., Li, X., and Duvenaud, D. (2018). Inference suboptimality in variational autoencoders. In *International conference on machine learning*, pages 1078–1086. PMLR.

Fraccaro, M., Sønderby, S. K., Paquet, U., and Winther, O. (2016). Sequential neural models with stochastic layers. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 2207–2215, Red Hook, NY, USA. Curran Associates Inc.

Ganguly, A., Jain, S., and Watchareeruetai, U. (2023). Amortized variational inference: A systematic review. *Journal of Artificial Intelligence Research*, 78:167–215.

Geweke, J. and Tanizaki, H. (2001). Bayesian estimation of state-space models using the metropolis–hastings algorithm within gibbs sampling. *Computational statistics & data analysis*, 37(2):151–170.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *the Journal of machine Learning research*, 14(1):1303–1347.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., and Zhao, L. (2019). Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia tools and applications*, 78:15169–15211.

Kim, M. and Pavlovic, V. (2021). Reducing the amortization gap in variational autoencoders: A bayesian random function approach. *arXiv preprint arXiv:2102.03151*.

Kingma, D. P. and Welling, M. (2013). Auto-Encoding Variational Bayes. *arXiv e-prints*, page arXiv:1312.6114.

Kipf, T. and Welling, M. (2016). Variational graph auto-encoders. *NeurIPS Workshop on Bayesian Deep Learning (NeurIPS BDL)*, abs/1611.07308.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Krishnan, R., Liang, D., and Hoffman, M. (2018). On the challenges of learning with inference networks on sparse, high-dimensional data. In Storkey, A. and Perez-Cruz, F., editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 143–151. PMLR.

Margossian, C. C. and Blei, D. M. (2023). Amortized variational inference: when and why? *arXiv preprint arXiv:2307.11018*.

Margossian, C. C. and Blei, D. M. (2024). Amortized variational inference: When and why? In *Uncertainty in Artificial Intelligence*, pages 2434–2449. PMLR.

Nasios, N. and Bors, A. G. (2006). Variational learning for gaussian mixture models. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36(4):849–862.

Rezende, D. and Mohamed, S. (2015). Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR.

Roberts, S. J., Husmeier, D., Rezek, I., and Penny, W. (1998). Bayesian approaches to gaussian mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1133–1142.

Tomczak, J. and Welling, M. (2018). Vae with a vampprior. In *International conference on artificial intelligence and statistics*, pages 1214–1223. PMLR.

Van Erven, T. and Harremos, P. (2014). Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820.

Wainwright, M. J., Jordan, M. I., et al. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305.

Wang, H., Bhattacharya, A., Pati, D., and Yang, Y. (2022). Structured variational inference in bayesian state-space models. In *International Conference on Artificial Intelligence and Statistics*, pages 8884–8905. PMLR.

West, M. and Harrison, J. (2006). *Bayesian forecasting and dynamic models*. Springer Science & Business Media.

Xing, E. P., Jordan, M. I., and Russell, S. (2012). A generalized mean field algorithm for variational inference in exponential families. *arXiv preprint arXiv:1212.2512*.

Yang, Y., Pati, D., and Bhattacharya, A. (2020). α-variational inference with statistical guarantees. *The Annals of Statistics*, 48(2):886–905.

Zeng, Y. (2013). *State-Space Models applications in economics and finance*. Springer.

Zhang, C., Bütepage, J., Kjellström, H., and Mandt, S. (2018). Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026.

Zhang, M., Hayes, P., and Barber, D. (2022). Generalization gap in amortized inference. *Advances in neural information processing systems*, 35:26777–26790.

Zhao, Y. and Linderman, S. (2023). Revisiting structured variational autoencoders. In *International Conference on Machine Learning*, pages 42046–42057. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]
   In Section 4, we listed 1.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
   We put our theorem in section 3.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) Code repository URL: `https://github.com/waterism211/Amortized-Structured-\Variational-Inference`.
   (c) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (d) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (e) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Yes]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Amortized Structural Variational Inference
Supplementary Materials

Code repository:
https://github.com/waterism211/Amortized-Structured-Variational-Inference.

1 Fully Connected Neural Network Architecture

In this section, we provide a comprehensive overview of the architecture of fully connected neural networks. The architecture of a neural network, denoted by $(L, \mathbf{k})$, is characterized by two primary components: the number of hidden layers L, which is a positive integer, and the width vector $\mathbf{k} = (k_1, \ldots, k_L) \in \mathbb{N}^L$, which specifies the number of neurons in each of the L hidden layers.

A multilayer feedforward neural network with architecture $(L, \mathbf{k})$, employing the ReLU activation function $\rho(x) = \max(0, x)$ for any $x \in \mathbb{R}$, can be mathematically represented as a real-valued function $f : \mathbb{R}^d \to \mathbb{R}$ defined as

$$f(\mathbf{x}) = \sum_{i=1}^{k_L} c_{1,i}^{(L)} f_i^{(L)}(\mathbf{x}) + c_{1,0}^{(L)}$$

for some weights $c_{1,0}^{(L)}, \ldots, c_{1,k_L}^{(L)} \in \mathbb{R}$, and for $f_i^{(L)}$ recursively defined by

$$f_i^{(s)}(\mathbf{x}) = \rho\left(\sum_{j=1}^{k_{s-1}} c_{i,j}^{(s-1)} f_j^{(s-1)}(\mathbf{x}) + c_{i,0}^{(s-1)}\right)$$

for some $c_{i,0}^{(s-1)}, \ldots, c_{i,k_{s-1}}^{(s-1)} \in \mathbb{R}$, for $s \in \{2, \ldots, L\}$, and

$$f_i^{(1)}(\mathbf{x}) = \rho\left(\sum_{j=1}^{d} c_{i,j}^{(0)} x^{(j)} + c_{i,0}^{(0)}\right)$$

for some $c_{i,0}^{(0)}, \ldots, c_{i,d}^{(0)} \in \mathbb{R}$.

For simplification, we assume that all hidden layers possess an identical number of neurons. Consequently, we define the space $\mathcal{F}(L, r)$, as per Kohler and Langer (2021), to represent the set of neural networks with L hidden layers and r neurons per layer:

Definition 1.1 (Neural Network Space) *The space of neural networks with L hidden layers and r neurons per layer is defined by*

$$\mathcal{F}(L, r) = \{f : f \text{ is of the form (4) and (5) with } k_1 = k_2 = \ldots = k_L = r\}.$$

Here, the neurons in the neural network can be viewed as computational units. The input of any neuron i in the s-th layer is associated with the output of all k_{s-1} units j in the $(s-1)$-th layer with weights $c_{i,j}^{(s-1)}$ through the ReLU function $\rho(\cdot)$. Such a kind of recursive construction of a fully connected neural network can be represented as an acyclic graph.

One key feature of our neural network architecture is that no network sparsity assumption is needed. The network class $\mathcal{F}(L, r)$ represents a fully connected feedforward neural network, where each neuron is connected to every neuron in the previous layer. To avoid the so-called curse of dimensionality, we consider f^* to be in the space of hierarchical composition models $\mathcal{H}(l, \mathcal{P})$ for some $l \in \mathbb{Z}^+$ and $\mathcal{P} \subseteq [1, \infty) \times \mathbb{Z}^+$, which is described next.

Definition 1.2 ((p, C)-Smooth Function) *Let $p = q + s$ for some $q \in \mathbb{N} = \mathbb{Z}^+ \cup \{0\}$ and $0 < s \leq 1$. A function $g : \mathbb{R}^d \to \mathbb{R}$ is called (p, C)-smooth if for every $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$ with $\|\boldsymbol{\alpha}\|_1 = q$, the partial derivative*

$$\frac{\partial^q g}{\partial z_1^{\alpha_1} \ldots \partial z_d^{\alpha_d}}$$

exists and satisfies

$$\left| \frac{\partial^q g}{\partial z_1^{\alpha_1} \ldots \partial z_d^{\alpha_d}}(\mathbf{z}) - \frac{\partial^q g}{\partial z_1^{\alpha_1} \ldots \partial z_d^{\alpha_d}}(\mathbf{w}) \right| \leq C \|\mathbf{z} - \mathbf{w}\|^s$$

for all $\mathbf{z}, \mathbf{w} \in \mathbb{R}^d$, where $\|\cdot\|$ denotes the Euclidean norm.

Definition 1.3 (Space of Hierarchical Composition Models, Kohler and Langer (2021)) *Let $l \in \mathbb{Z}^+$ and $\mathcal{P} \subseteq [1, \infty) \times \mathbb{Z}^+$. The space of hierarchical composition models is defined recursively.*

For $l = 1$,

$$\mathcal{H}(1, \mathcal{P}) := \Big\{ h : \mathbb{R}^d \to \mathbb{R} : h(\mathbf{z}) = u(z^{(\pi(1))}, \ldots, z^{(\pi(K))}), \\ \text{where } u : \mathbb{R}^K \to \mathbb{R} \text{ is } (p, C)\text{-smooth for some } (p, K) \in \mathcal{P}, \text{ and } \pi : \{1, \ldots, K\} \to \{1, \ldots, d\} \Big\}.$$

For $l > 1$, we recursively define

$$\mathcal{H}(l, \mathcal{P}) := \big\{ h : \mathbb{R}^d \to \mathbb{R} : h(\mathbf{z}) = u(f_1(\mathbf{z}), \ldots, f_K(\mathbf{z})), \\ \text{where } u : \mathbb{R}^K \to \mathbb{R} \text{ is } (p, C)\text{-smooth for some } (p, K) \in \mathcal{P}, \text{ and } f_i \in \mathcal{H}(l-1, \mathcal{P}) \big\}.$$

Here, the function class $\mathcal{H}(l, \mathcal{P})$ describes the relationships between the input and output of the network, where $(p, K) \in \mathcal{P}$ describes the smoothness and order constraint of the hierarchical composition model. Notably, additive models, single index models, and projection pursuit can be viewed as special cases of the hierarchical composition model.

2 Appendix B: The α-Variational Inference Framework

In this appendix, we provide background on the α-variational inference (VI) framework that underpins the theoretical results in Section 3. We begin by reviewing the classical variational formulation, then motivate and define the α-VI objective, and conclude with a comparison to the standard VI objective.

Classical variational inference We consider a parametric latent variable model with parameter $\theta = (\mu, \lambda)$, observed data X^n, and latent variables Z^n. The marginal likelihood under parameter θ is

$$p_\theta^{(n)}(X^n) = \int p(X^n \mid Z^n, \mu)\, p(Z^n \mid \lambda)\, dZ^n.$$

The true data-generating distribution is denoted by $p_{\theta^*}^{(n)}$, where $\theta^* = (\mu^*, \lambda^*)$ is the true parameter. The marginal log-likelihood ratio is defined as

$$\ell_n(\theta, \theta^*) := \log p_\theta^{(n)}(X^n) - \log p_{\theta^*}^{(n)}(X^n).$$

In standard variational inference, we approximate the posterior $p(Z^n, \theta \mid X^n)$ using a structured variational family of the form

$$q(Z^n, \theta) = q_{\gamma(X^n)}(Z^n) \cdot q_\theta(\theta),$$

where the latent component is amortized via a deterministic encoder γ. The variational distribution is chosen to minimize the KL divergence:

$$D\left[q_{\gamma(X^n)}(Z^n) \cdot q_\theta(\theta) \,\|\, p(Z^n, \theta \mid X^n) \right].$$

This is equivalent to maximizing the evidence lower bound (ELBO), which admits the decomposition:

$$\log p_\theta^{(n)}(X^n) = \text{ELBO}(q) + D\left[q \,\|\, p(Z^n, \theta \mid X^n) \right],$$

so that minimizing the KL divergence corresponds to maximizing a lower bound on the marginal log-likelihood.

However, in models with latent variables, directly analyzing the ELBO becomes difficult due to intractable expectations and approximation artifacts. In particular, the ELBO may include a Jensen gap due to the variational approximation over latent variables.

A reformulation of the variational objective To better understand the structure of the ELBO in latent variable models, we rewrite the variational objective as

$$\Psi_n(q_\theta, q_{\gamma(X^n)}) := -\int_\Theta \ell_n(\theta, \theta^*)\, q_\theta(d\theta) + \Delta_J(q_\theta, q_{\gamma(X^n)}) + D(q_\theta \,\|\, p_\theta),$$

where the Jensen gap Δ_J quantifies the error due to the latent variational approximation:

$$\Delta_J(q_\theta, q_{\gamma(X^n)}) := \int_\Theta \left[\ell_n(\theta) - \widehat{\ell}_n^{\gamma}(\theta)\right] q_\theta(d\theta),$$

and $\widehat{\ell}_n^{\gamma}(\theta)$ is the amortized surrogate log-likelihood:

$$\widehat{\ell}_n^{\gamma}(\theta) := \int q_{\gamma(X^n)}(Z^n) \log\left(\frac{p(X^n \mid Z^n, \mu) \cdot p(Z^n \mid \lambda)}{q_{\gamma(X^n)}(Z^n)}\right) dZ^n.$$

Minimizing Ψ_n is equivalent to minimizing the original KL divergence to the posterior, up to an additive constant $\ell_n(\theta^*)$. This decomposition separates the model-fit components (log-likelihood ratio and Jensen gap) from the regularization term, and motivates the next extension.

The α-VI objective To address challenges in theoretical analysis and improve control over approximation error, the α-VI framework introduces an inverse temperature parameter $\alpha \in (0, 1]$. The associated objective is defined as:

$$\Psi_{n,\alpha}(q_\theta, q_{\gamma(X^n)}) = -\int_\Theta \ell_n(\theta, \theta^*)\, q_\theta(d\theta) + \Delta_J(q_\theta, q_{\gamma(X^n)}) + \frac{1}{\alpha} D(q_\theta \,\|\, p_\theta).$$

As α decreases, the regularization term is upweighted, providing stronger control over overfitting and facilitating concentration guarantees. In particular, for $\alpha = 1$, this recovers the classical VI objective Ψ_n above.

Interpretation and summary The α-VI objective maintains the same structure as standard VI but introduces an explicit regularization trade-off via α. This results in more tractable analysis under mild conditions. The solution to the α-VI problem is given by

$$(\widehat{q}_{\theta,\alpha}, \widehat{q}_{\gamma(X^n),\alpha}) := \arg \min_{q_\theta \in \Gamma_\theta,\, q_{\gamma(X^n)} \in \mathcal{Q}_{\text{AS}}} \Psi_{n,\alpha}(q_\theta, q_{\gamma(X^n)}),$$

which is used throughout the theoretical developments in Section 3 to derive non-asymptotic variational risk bounds.

Framework follows prior α-VI and structured-VI analyses, with modifications to accommodate amortized, structured variational distributions over Z^n via encoder mappings γ.

3 Additional Simulation Studies

In this section, we conduct additional simulation studies by generating data from several alternative models and comparing our proposed methods with existing approaches.

3.1 Algorithm of Amortized Neighborhood variational Inference(ANVI)

The ANVI algorithm as described in Algorithm 1 is implemented analogously to the ASVI algorithm presented in the main paper; the only difference in ANVI is that we do not perform conditional sampling.

3.2 Hidden Markov Model with more learnable parameters

Consider the same model as we discussed in the simulation part. We add one more global parameter A here. Firstly, we consider X^n only depends on Z_i linearly:

$$Z_i = A \cdot Z_{i-1} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \tau^2), \quad X_i = \theta + Z_i + \eta_i, \quad \eta_i \sim \mathcal{N}(0, \sigma^2). \tag{1}$$

In the simulation setting, we will keep the sample size $n = 500$ and vary A. We will compare the metric in Appendix 3.6. We can see in Table 1 our ASVI out performs the MFVI and gets very close to SVI.

Algorithm 1 Amortized Neighborhood-Aware Variational Inference (ANVI) Optimization

Input: Data $\{X_1, X_2, \ldots, X_n\}$, contextualized as $X_C^n = \{X_{C_1}, \ldots, X_{C_n}\}$ with neighborhood information.
Output: Variational parameters $\{\mu_z, \Sigma_z, \mu_\theta, \Sigma_\theta\}$, variational approximation $q(z, \theta) = \mathcal{N}(\mu_z, \Sigma_z)\mathcal{N}(\mu_\theta, \Sigma_\theta)$
Initialize $\{\mu_z, \Sigma_z, \mu_\theta, \Sigma_\theta\}$
while the loss $\mathcal{L}(X^n; \theta, Z^n)$ has not converged **do**
    Sample $\{\theta_j\}_{j=1}^m$ from $\mathcal{N}(\mu_\theta, \Sigma_\theta)$
    for $i = 1$ to n **do**
        Sample $\{Z_{ij}\}_{j=1}^m$ from $\mathcal{N}(\mu_z(X_{C_i}), \Sigma_z(X_{C_i}))$
        Compute $\mathcal{L}(X_i; \theta, Z_i) = \sum_{j=1}^m [\log p_\alpha(X_i, Z_{ij}, \theta_j) - \log q(\theta_j, Z_{ij} \mid X_i)]$
        Compute $\nabla \log \mathcal{L}(X_i; \theta, Z_i)$ w.r.t. $\{\mu_z, \Sigma_z, \mu_\theta, \Sigma_\theta\}$
    end for
    $\widehat{\nabla} \log \mathcal{L} = \frac{1}{n} \sum_{i=1}^n \nabla \log \mathcal{L}(X_i; \theta_i, Z_i)$
    Update $\{\mu_z, \Sigma_z, \mu_{\theta_i}, \Sigma_\theta\}$ using $\widehat{\nabla} \log \mathcal{L}$
end while
Return: $\{\mu_z, \Sigma_z, \mu_\theta, \Sigma_\theta\}$

Table 1: Nonlinear AR(1) performance summary across A_{true}.

A_{true}	Method	ELBO	MSE (global)	MSE (pred)
0.30	MFVI	-240	0.0310	0.3358
	SVI	**-225**	**0.0031**	**0.3264**
	ASVI	-235	0.0038	0.3295
0.50	MFVI	-245	0.0693	0.3488
	SVI	**-225**	**0.0051**	**0.3232**
	ASVI	-235	0.0116	0.3289
0.60	MFVI	-255	0.0677	0.3492
	SVI	**-235**	**0.0120**	**0.3199**
	ASVI	-245	0.0246	0.3268
0.70	MFVI	-290	0.0677	0.3492
	SVI	**-255**	**0.0120**	**0.3199**
	ASVI	-270	0.0246	0.3268
0.80	MFVI	-320	0.0382	0.3240
	SVI	**-255**	**0.0241**	**0.3192**
	ASVI	-285	0.0380	0.3209

3.3 Higher-Order Hidden Markov Model

We generate data from the following second-order hidden Markov model:

$$Z_i = 0.6\, Z_{i-1} + 0.4\, Z_{i-2} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \tau^2), \tag{2}$$

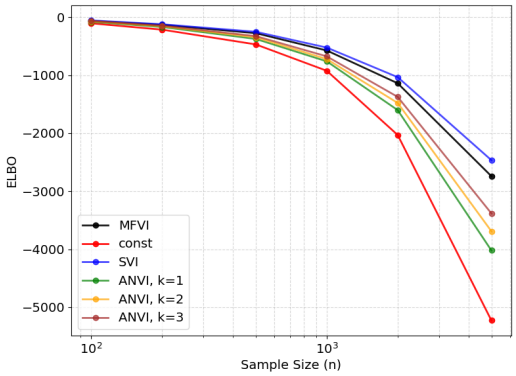
$$X_i = \theta + \sin(Z_i) + \eta_i, \quad \eta_i \sim \mathcal{N}(0, \sigma^2), \tag{3}$$

where $\theta = 2$, $\tau = 0.5$, and $\sigma = 0.7$. We simulate datasets of various sample sizes under this specification.

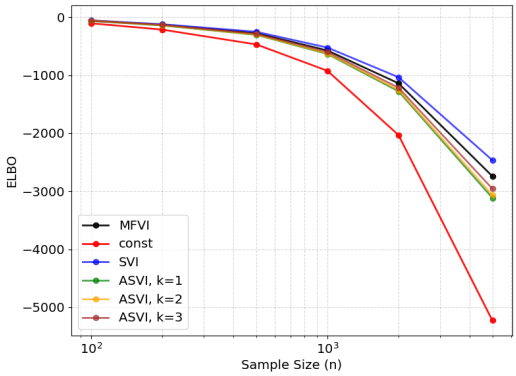
Figure 1 presents the final ELBO comparison between our proposed methods and the standard approaches. As shown in Figure 1a, incorporating neighborhood information helps reduce the amortization gap. When $k = 3$, using the true neighborhood size yields the greatest reduction in this gap. In contrast, for ASVI (Figure 1b), since the true model is not a first-order HMM, ASVI no longer outperforms MFVI. If our ASVI uses $\mathcal{Q}_{\text{AS}} = \{q : q(\theta, Z^n) = q_0(\theta) q_{\gamma(X_1)}(Z_1) q_{\gamma(X_2)}(Z_2|Z_1) \prod_{i=3}^n q_{\gamma(X_i)}(Z_i|Z_{i-1}, Z_{i-2})\}$, we can still out perform MFVI. The SVI works well since it has the flexibility to train each conditional posterior. Nonetheless, including neighborhood information still narrows the gap. In Figure 2, we still observe that our amortized methods converge in significantly fewer iterations.

3.4 Graphical Hidden Model

We consider a model with local hidden variables defined on a known underlying graph. Let the latent variables follow a multivariate normal distribution governed by the graph's adjacency matrix A, where each node has three neighbors.

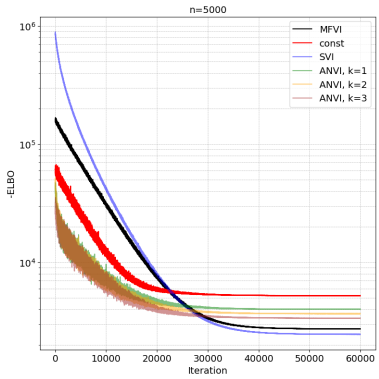


(a) ELBO of the ANVI methods

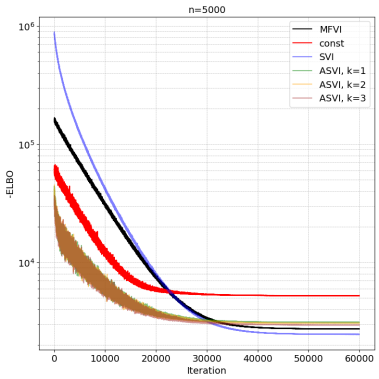


(b) ELBO of the ASVI methods

Figure 1: The ELBO for different VI methods with sample size. Here k stands for the number of neighbors for the amortization map. A large ELBO is preferred.



(a) ANVI



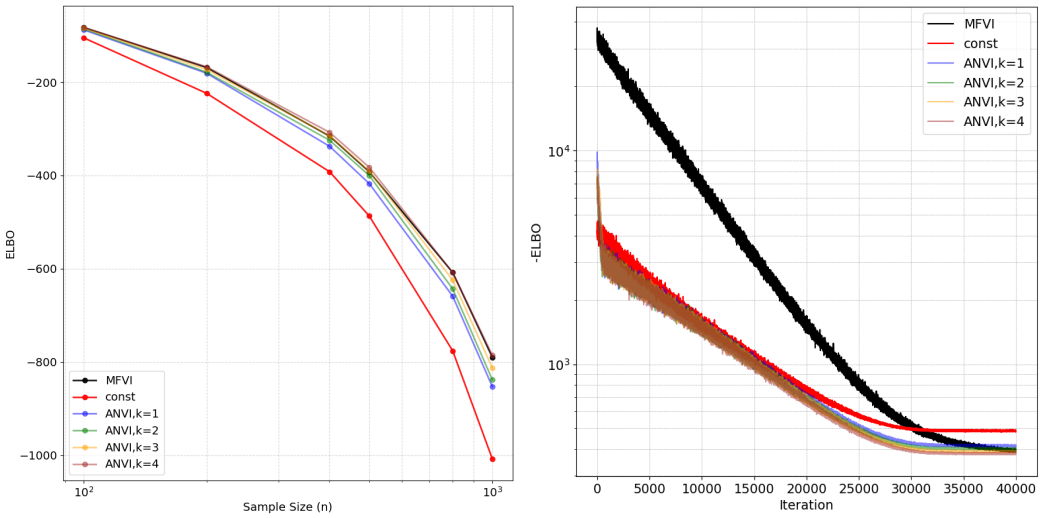
(b) ASVI

Figure 2: The optimization sample paths of different methods. A small value at the final iteration is preferred.

Specifically, we simulate data from

$$
\begin{aligned}
p(\theta) &\propto 1, \\
Z_{1:n} &\sim \mathcal{N}\big(0, \tau^2\big(I_n + \tfrac{1}{3}A\big)\big), \\
X_i &= \theta + \sin(Z_i) + \eta_i, \quad \eta_i \sim \mathcal{N}(0, \sigma^2).
\end{aligned}
$$

We then apply ANVI to this model to assess how incorporating neighborhood information narrows the amortization gap. As shown in Figure 3a, using the true neighborhood specification drives the gap nearly to zero.



(a) The ELBO for different variational family and sample size.

(b) The optimization sample paths of structured VI family and standard VI family.

Figure 3: Results for Graphical Hidden Model

3.5 The sensitivity of MLP

In this section, we explore the impact of different neural network architectures on the hidden Markov Model introduced in the main paper. Specifically, we evaluate the evidence lower bound (ELBO) under varying network widths. As illustrated in Figure 4, the ELBO values remain consistent across all configurations, indicating that our method is robust to changes in network width.

3.6 implementation detail for AR(p) model and nonlinear AR(2) model

3.6.1 AR(p) model

We consider a 3–dimensional latent AR(64) process $\{Z_i\}_{i=1}^n$ with an in simulation, and AR(3) observations $\{X_i\}_{i=1}^N$ in Section 4.3:

$$
Z_i = \sum_{j=1}^{3} a_j \, Z_{i-j} \; + \; \sum_{k=4}^{64} c_k \, Z_{i-k} \; + \; \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \tau^2 I_3), \tag{4}
$$

$$
X_i = \theta \; + \; Z_i \; + \; \sum_{j=1}^{3} b_j \, X_{i-j} \; + \; \eta_i, \qquad \eta_i \sim \mathcal{N}(0, \sigma^2 I_3). \tag{5}
$$

The AR(64) tail coefficients in simulation are $c_k = \frac{1}{10k^2}$ for $k \geq 4$, with an optional stability rescaling $s \in (0, 1]$ so that $\sum_{j=1}^{3} |a_j| + \sum_{k \geq 4} |s\, c_k| \leq \rho_{\text{sim}} < 1$; setting $\gamma = 0$ removes the tail. The joint density factorizes as $p_\theta(X_{1:n}, Z_{1:n}) = \prod_{i=1}^{n} p(Z_i \mid Z_{1:i-1})\; p(X_i \mid X_{1:i-1}, Z_i)$, with $p(Z_i \mid \cdot)$ and $p(X_i \mid \cdot)$ given by (4)–(5).

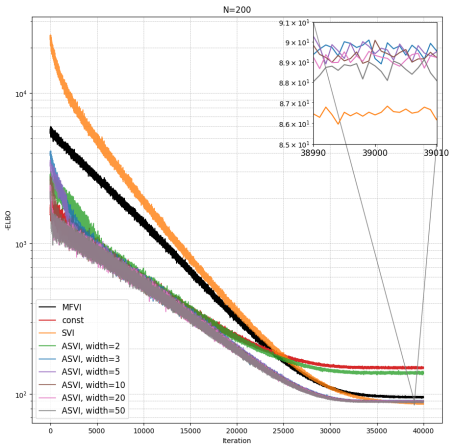


Figure 4: The sample path for ASVI with the different neural network structure, a larger ELBO value is preferred as better method. Here width stands for the number of neuron in the neural network structure. The zoom in part is trying to distinguish these methods with similar ELBO value.

We place Gaussian priors on a global offset θ and on PACF latents for both $\mathbf{a}$ and $\mathbf{b}$:

$$\theta \sim \mathcal{N}(0, \sigma_\theta^2), \quad u_{a,j} \sim \mathcal{N}(\mu_{a,j}, s_{a,j}^2), \quad u_{b,j} \sim \mathcal{N}(\mu_{b,j}, s_{b,j}^2),$$

and map $u \mapsto$ PACF via $k_j = \tanh(u_{a,j})$, $r_j = \tanh(u_{b,j})$ to keep $k_j, r_j \in (-1, 1)$. The AR(3) coefficients are then obtained by the Durbin–Levinson recursion $\mathcal{T}(\cdot)$:

$$(a_1, a_2, a_3) = \mathcal{T}(k_1, k_2, k_3), \qquad (b_1, b_2, b_3) = \mathcal{T}(r_1, r_2, r_3),$$

which guarantees a stable AR(3).

We approximate the posterior with $q(\theta)\, q(u_a)\, q(u_b)\, q(Z_{1:n} \mid X_{1:n})$ and optimize the *true* ELBO:

$$\mathcal{L} = \mathbb{E}_q[\log p_\theta(X_{1:n}, Z_{1:n}) + \log p(\theta) + \log p(u_a) + \log p(u_b) - \log q(\theta) - \log q(u_a) - \log q(u_b) - \log q(Z_{1:n} \,|\, X_{1:n})]\,.$$

We use $q(\theta), q(u_a), q(u_b)$ as univariate Gaussians (reparameterized). For $q(Z_{1:n} \,|\, X_{1:n})$ we consider:

- **MFVI:** factorized Gaussian, $q_{\text{MF}}(Z_{1:n} \,|\, X_{1:n}) = \prod_{i=1}^n \mathcal{N}(Z_i \mid \mu_i, \text{diag}(s_i^2))$.
- **SVI (structured, non-amortized):** a linear–Gaussian AR(3) sampler via reparameterization

  $$\begin{aligned} Z_1 &\sim \mathcal{N}(a^{(1)}, s_1^2 I), \\ Z_2 &\sim \mathcal{N}(a^{(2)} + b_1^{(2)} Z_1,\ s_2^2 I), \\ Z_i &\sim \mathcal{N}\big(a_i + b_{1,i-1} Z_{i-1} + b_{2,i-2} Z_{i-2} + b_{3,i-3} Z_{i-3},\ s_i^2 I\big), \quad i \geq 3, \end{aligned}$$

  where all $\{a_i, b_{1,i-1}, b_{2,i-2}, b_{3,i-3}, s_i\}$ are free per-index variational parameters.
- **ASVI (amortized, structured):** the same conditional Gaussian AR(3) form as SVI, but the per-index parameters are *deterministic functions* of X produced by an encoder g_ϕ with a sliding window of length k:

  $$(a_i, b_{1,i-1}, b_{2,i-2}, b_{3,i-3}, s_i) = g_\phi\big(X_{i:i+k}\big).$$

- **FlowASVI:** draw $Z_{1:n}^{(0)} \sim q_{\text{ASVI}}(\cdot \mid X_{1:n})$ and apply a conditional normalizing flow $Z_{1:n} = f_\psi(Z_n^{(0)};\, X_{1:n})$ (affine RealNVP couplings with context from an MLP over windows of X). The density corrects as

  $$\log q_{\text{flow}}(Z_{1:n} \,|\, X_{1:n}) = \log q_{\text{ASVI}}(Z_{1:n}^{(0)} \,|\, X_{1:n}) - \sum_\ell \log\big|\det \tfrac{\partial f_\psi^{(\ell)}}{\partial Z^{(\ell-1)}}\big|.$$

  Each coupling layer uses a binary mask $m \in \{0,1\}^D$ and context c_i; componentwise

  $$Y = m \odot Z + (1 - m) \odot \big(Z \odot e^{s(c)} + t(c)\big), \quad \log|\det J| = \sum_{j : m_j = 0} s_j(c),$$

All expectations are estimated with reparameterized Monte Carlo (antithetic pairs) and optimized with AdamW. We train in two stages (temporarily freezing b–heads) and add mild smoothness/centering regularizers on structured parameters.

For any trained method, we report

$$\widehat{\mathrm{ELBO}} = \frac{1}{B}\sum_{b=1}^{B}\Big\{\log p_\theta(X_{1:n}, Z_{1:n}^{(b)}) + \log p(\theta) + \log p(u_a) + \log p(u_b) - \log q(\theta) - \log q(u_a) - \log q(u_b)\Big\} - \log q(Z_{1:n}^{(b)} \mid X_{1:n}).$$

Using the posterior mean path $\bar{Z}_i = \mathbb{E}_q[Z_i \mid X_{1:n}]$, we form the conditional mean predictor with teacher forcing:

$$\widehat{X}_i = \theta + \bar{Z}_i + \sum_{j=1}^{3} b_j X_{i-j}, \qquad \mathrm{MSE} = \frac{1}{3n}\sum_{i=1}^{n}\|X_i - \widehat{X}_i\|_2^2.$$

For the AR(3) model without model misspecification:

Table 2: AR(3) performance summary

Method	ELBO	MSE (global)	MSE (pred)	Time (s)
MFVI	-873.59	0.2252	0.3272	27.3
SVI	**-811.13**	**0.0082**	**0.3147**	1272.1
ASVI_k1	-834.41	0.0305	0.3588	1237.2
ASVI_k2	-838.57	0.0281	0.3399	1241.3
ASVI_k3	-816.76	0.0324	0.3238	1245.1
FlowASVI_k1	-841.98	0.0960	0.3685	1270.5
FlowASVI_k2	-835.50	0.1024	0.3564	1267.9
FlowASVI_k3	-837.18	0.1040	0.3387	1245.5

3.6.2 Non-linear AR(2) model

For ASVI, we use an amortized variational guide that conditions on the local context and the two latent lags:

$$\mathrm{context}_i \;=\; \big[X_{i-2},\, X_{i-1},\, X_i,\; Z_{i-2},\, Z_{i-1}\big],$$

The guide is a bounded MLP that outputs a diagonal Gaussian:

$$q_{\phi(\mathrm{context}_i)}(Z_i \mid Z_{i-1}, Z_{i-2}) \;=\; \mathcal{N}\Big(Z_i;\; \mu_i,\; \mathrm{diag}(\sigma_i^2)\Big), \qquad \begin{aligned} h_i &= \mathrm{MLP}_\phi(\mathrm{context}_i),\\ \mu_i &= Z_{i-1} \;+\; \tanh(W_\mu h_i)\,\cdot s_\mu,\\ \log\sigma_i^2 &= 2\log\big(\mathrm{softplus}(\underbrace{\tanh(W_s h_i)\, s_s}_{\text{bounded}})\big) + \log\tau, \end{aligned}$$

where s_μ and s_s are scalar caps that keep the residual update and the log-scale bounded, and $\tau > 0$ is a learned temperature (softplus reparametrized) to calibrate posterior variance. Compared to a per-time-step linear SVI (with affine dependence on (z_{i-1}, z_{i-2})), the bounded MLP in ASVI captures richer nonlinearities in the posterior while remaining numerically stable.

For flow-ASVI, to further enrich the posterior family, we also consider a flow-based guide that transforms a diagonal-Gaussian base via K planar flows. Given the same context $h_i = \mathrm{MLP}_\phi(\mathrm{context}_i)$, we predict the base parameters and flow parameters:

$$Z_0 \sim \mathcal{N}\big(\mu_i,\, \mathrm{diag}(\sigma_i^2)\big), \qquad \{w_k(h_i), u_k(h_i), b_k(h_i)\}_{k=1}^{K}.$$

Each planar layer applies

$$Z_k \;=\; Z_{k-1} \;+\; \hat{u}_k\, \tanh\big(w_k^\top Z_{k-1} + b_k\big), \quad \log\Big|\det\frac{\partial Z_k}{\partial Z_{k-1}}\Big| = \log\Big|1 + \hat{u}_k^\top \psi_k(Z_{k-1})\Big|,$$

with the standard stabilization $\hat{u}_k = u_k + \frac{(m - w_k^\top u_k)}{\|w_k\|^2} w_k$ and $\psi_k(Z) = (1 - \tanh^2(w_k^\top Z + b_k))\, w_k$, $m = -1 + \mathrm{softplus}(w_k^\top u_k)$.

Trade-offs vs. ASVI. While Flow-ASVI may occasionally nudge the ELBO upward due to its higher expressivity, it is (i) heavier to train (additional flow parameters and log-determinant costs), (ii) more sensitive to optimization (flow instabilities, parameter collapse), and (iii) in our experiments often produced worse downstream *latent* RMSE and larger global MSE than ASVI. In contrast, the bounded-MLP ASVI achieves a strong balance of flexibility, stability, and efficiency, resulting in more reliable latent recovery and global parameter estimation in practice.

3.7 The Moving MNIST data for reconstruction and prediction

Moving MNIST dataset is proposed by Srivastava et al. (2015). It consists of short grayscale clips (usually 20 frames of size 64×64) in which one or two hand-written MNIST digits move linearly across the frame and "bounce" off the boundaries. The data are stored as a NumPy array of shape (T, N, 64, 64)—T time steps and N independent sequences. Because the digits' motions are random but rule-based, Moving MNIST offers a controlled yet non-trivial testbed for models that learn spatial structure. In our simulation study, we pick the odds number of these sequence(10 frames of size 64×64).

We consider Moving-MNIST sequences $X_{1:T}$, where each frame $X_i \in [0,1]^{64\times 64}$ is a single grayscale image. Unless otherwise noted, all models operate per frame or per (short) window.

We learn a latent representation $Z_i \in \mathbb{R}^d$ (d=64) via an encoder–decoder pair

$$q_\phi(Z_i \mid X_i) = \mathcal{N}\big(\mu_\phi(X_i), \mathrm{diag}(\sigma_\phi^2(X_i))\big)\,, \tag{6}$$

$$\hat{X}_i = f_\theta(Z_i), \tag{7}$$

where μ_ϕ, σ_ϕ are produced by a convolutional encoder (GroupNorm, ReLU), and f_θ is a deconvolutional decoder (GroupNorm, ReLU, Sigmoid). We train with a β-VAE loss and "free-bits" regularization:

$$\mathcal{L}_{\text{VAE}}(\theta,\phi) = \mathbb{E}_{X\sim\mathcal{D}}\left[\underbrace{\ell_{\text{rec}}(X,\hat{X})}_{\text{reconstruction}} + \beta\cdot\underbrace{\mathrm{KL}_\tau(q_\phi(Z\,|\,X)\,\|\,\mathcal{N}(0,I))}_{\text{KL with free-bits}}\right], \tag{8}$$

$$\mathrm{KL}_\tau = \sum_{i=1}^{d}\max\{\mathrm{KL}_i,\ \tau\}, \tag{9}$$

with ℓ_{rec} taken as per-pixel MSE in our runs. We use KL warm-up to a final β and a small free-bits floor τ.

For dynamics we work in a whitened latent space. With $\hat{\mu}, \hat{\sigma}$ the per-dimension mean/std of $\mu_\phi(X)$ computed on train data,

$$Z_i^{\text{w}} = \frac{\mu_\phi(X_i) - \hat{\mu}}{\hat{\sigma}}, \qquad \text{and} \qquad \mathrm{inv}(Z^{\text{w}}) = Z^{\text{w}}\odot\hat{\sigma} + \hat{\mu}. \tag{10}$$

3.7.1 One-step priors in latent space

Given context c_i and target Z_{i+1}^{w}, we train four priors to predict the next whitened latent.

Random baseline.

$$p_{\text{rand}}(Z_{i+1}^{\text{w}}) = \mathcal{N}(0, I).$$

ASVI. Given a window of whitened latents $C_i = \big[Z_{i-w+1}^{\text{w}}, \ldots, Z_i^{\text{w}}\big] \in \mathbb{R}^{w\times d}$, our *test-time* prior is a windowed MLP

$$\hat{Z}_{i+1,p}^{\text{w}} \;=\; f_\varphi\big(\,\mathrm{vec}(C_i)\big). \tag{11}$$

During training only, we introduce a *guide* that may peek at the next image,

$$\hat{Z}_{i+1,q}^{\text{w}} \;=\; g_\gamma\big(\,\mathrm{vec}(C_i),\, X_{i+1}\big), \tag{12}$$

and *distill* its signal into the prior. The total loss is

$$\mathcal{L}_{\text{ASVI}}(\varphi,\gamma) = \underbrace{\big\|\hat{Z}_{i+1,p}^{\text{w}} - Z_{i+1}^{\text{w}}\big\|_2^2}_{\text{prior MSE}} + \lambda_{\text{distill}}\underbrace{\big\|\hat{Z}_{i+1,p}^{\text{w}} - \mathrm{sg}\big(\hat{Z}_{i+1,q}^{\text{w}}\big)\big\|_2^2}_{\text{distill to guide}} + \lambda_{\text{guide}}\underbrace{\big\|\hat{Z}_{i+1,q}^{\text{w}} - Z_{i+1}^{\text{w}}\big\|_2^2}_{\text{guide MSE}}, \tag{13}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. **At evaluation** we discard the guide and use only f_φ to predict $\hat{Z}_{i+1}^{\text{w}}$.

Gaussian. To match ASVI's context, we use a window-conditioned Gaussian mean regressor

$$(\mu_\psi, \ell_\psi) \;=\; h_\psi\big(\,\mathrm{vec}(C_i)\big), \qquad p_\psi\big(Z^{\mathrm{w}}_{i+1} \mid C_i\big) = \mathcal{N}\big(\mu_\psi,\; \mathrm{diag}(\exp \ell_\psi)\big). \tag{14}$$

We train it via a residual (ΔZ) target using the last latent in the window as the base:

$$\hat{Z}^{\mathrm{w}}_{i+1} \;=\; Z^{\mathrm{w}}_i \;+\; \mu_\psi, \qquad \mathcal{L}_{\text{Gauss}}(\psi) \;=\; \big\|\hat{Z}^{\mathrm{w}}_{i+1} - Z^{\mathrm{w}}_{i+1}\big\|_2^2. \tag{15}$$

(Equivalently, h_ψ predicts ΔZ^{w}; we add it back to Z^{w}_i.) **No** access to X_{i+1} is used here.

Flow prior. Let the condition be $c_i = Z^{\mathrm{w}}_i$ (Markov) or $c_i = \mathrm{vec}(C_i)$ (window). We define an invertible mapping $g_\psi(\cdot; c_i)$ such that

$$\epsilon = g_\psi^{-1}\big(Z^{\mathrm{w}}_{i+1}; c_i\big)\,, \tag{16}$$

$$\epsilon \sim \mathcal{N}\big(\mu_0(c_i), \mathrm{diag}(\exp \ell_0(c_i))\big), \tag{17}$$

with affine coupling layers conditioned on c_i. We maximize conditional log-likelihood:

$$\mathcal{L}_{\text{Flow}}(\psi) = \mathbb{E}\big[-\log p_\psi(Z^{\mathrm{w}}_{i+1} \mid c_i)\big] = \mathbb{E}\left[\frac{1}{2}\|\epsilon - \mu_0(c_i)\|^2_{\mathrm{diag}(\exp \ell_0(c_i))^{-1}} + \frac{1}{2}1^\top \ell_0(c_i) - \log|\det J_{g_\psi}|\right]. \tag{18}$$

3.7.2 Prediction and evaluation

Deterministic one-step prediction. At test time we form $\hat{Z}^{\mathrm{w}}_{i+1}$ via the chosen prior (ASVI/Gauss/Flow mean; Random draws a single sample or MC-average for metrics), map back to the VAE space, and decode:

$$\hat{Z}_{i+1} = \mathrm{inv}\Big(\hat{Z}^{\mathrm{w}}_{i+1}\Big)\,, \qquad \hat{X}_{i+1} = f_\theta(\hat{Z}_{i+1}). \tag{19}$$

Metrics. We report per-frame image MSE:

$$\mathrm{MSE} = \frac{1}{N}\sum_{n=1}^{N} \|\hat{X}^{(n)} - X^{(n)}\|_2^2. \tag{20}$$

4 Proof of Theorem 1

We start by restating the theorem for completeness.

Theorem 4.1 *Let $\hat{q}_{\theta,\alpha}$ denote the α-Rényi variational posterior obtained from an amortized variational approximation $q_{W^n}(Z^n, \theta) = q_{\gamma(X^n)}(Z^n) \cdot q_\theta(\theta)$, where γ is a deterministic encoder depending on X^n. Let $\pi^*_{Z^n} := p(Z^n \mid X^n, \theta^*)$ denote the true latent posterior given θ^*. For any fixed $(\varepsilon_\lambda, \varepsilon_\mu) \in (0,1)^2$ and $D > 1$, with probability at least $1 - \frac{5}{(D-1)^2\left(f_\lambda(n)\varepsilon_\lambda^2 + n f_\mu(n)\varepsilon_\mu^2\right)}$ under the true parameter θ^*, where f_μ and f_λ are defined as*

$$\mathcal{B}^{\mathrm{VI}}_n(\lambda^*, \varepsilon_\lambda) := \left\{ \lambda \in \Lambda : \begin{array}{l} D\left(p(Z^n \mid \lambda^*) \,\|\, p(Z^n \mid \lambda)\right) \le f_\lambda(n)\varepsilon_\lambda^2 \\ V\left(p(Z^n \mid \lambda^*) \,\|\, p(Z^n \mid \lambda)\right) \le f_\lambda(n)\varepsilon_\lambda^2 \end{array} \right\}, \tag{21}$$

$$\mathcal{B}^{\mathrm{VI}}_n(\mu^*, \varepsilon_\mu) := \left\{ \mu \in \mathcal{M} : \begin{array}{l} \max\limits_{1\le i\le n} \mathbb{E}_{Z^n|\theta^*} D_i(\mu^*, \mu) \le f_\mu(n)\varepsilon_\mu^2 \\ \max\limits_{1\le i\le n} \mathbb{E}_{Z^n|\theta^*} V_i(\mu^*, \mu) \le f_\mu(n)\varepsilon_\mu^2 \end{array} \right\}, \tag{22}$$

with $D_i(\mu^*, \mu) := D\left[p(\cdot \mid \mu^*, X_i) \mid p(\cdot \mid \mu, X_i)\right]$, and $V_i(\mu^*, \mu) := V\left[p(\cdot \mid \mu^*, X_i) \mid p(\cdot \mid \mu, X_i)\right]$, it holds that

$$\begin{aligned} \int D^{(n)}_\alpha(\theta, \theta^*)\, \hat{q}_{\theta,\alpha}(d\theta) \le \frac{D\alpha}{1-\alpha}\left(\left[\frac{f_\lambda(n)}{n}\varepsilon_\lambda^2 + f_\mu(n)\varepsilon_\mu^2\right] + \left[\Delta^2_{\mathit{ASVIGap}}\right]\right) \\ - \frac{1}{n(1-\alpha)}\log P_\lambda\left(\mathcal{B}^{\mathrm{VI}}_n(\lambda^*, \varepsilon_\lambda)\right) - \frac{1}{n(1-\alpha)}\log P_\mu\left(\mathcal{B}^{\mathrm{VI}}_n(\mu^*, \varepsilon_\mu)\right). \end{aligned}$$

where

$$\inf_{q_{\gamma(X^n)} \in \mathcal{Q}_{AS}} \frac{1}{n} D\left(q_{\gamma(X^n)} \,\|\, \pi^*_{Z^n}\right) \le \Delta^2_{\mathit{ASVIGap}}. \tag{23}$$

Proof 1 *We begin by applying Theorem 3.1 from Yang et al. (2020), which provides a general PAC-Bayes-type upper bound for the α-Rényi variational risk. In our setting, we consider the amortized variational family defined by*

$$q_{\gamma(X^n)}(Z^n, \theta) := q_{\gamma(X^n)}(Z^n) \cdot q_\theta(\theta),$$

where the encoder map γ deterministically maps the data X^n to variational parameters for Z^n.

Let Γ_θ denote the collection of variational distributions over θ such that $q_\theta \ll p_\theta$, and let $\Gamma_{\gamma(X^n)}$ denote the class of amortized variational distributions over Z^n induced by encoder functions γ, i.e.,

$$\Gamma_{\gamma(X^n)} := \left\{ q(Z^n) = q_{\gamma(X^n)}(Z^n) \,:\, \gamma \in \mathcal{G} \right\},$$

where $\mathcal{G}$ is the space of admissible encoder mappings (e.g., neural networks).

We now apply Theorem 3.1 from Yang et al. (2020) with

$$\zeta := \frac{1}{(D-1)^2 \left(\varepsilon_\lambda^2 f_\lambda(n) + n \varepsilon_\mu^2 f_\mu(n) \right)},$$

which guarantees that with probability at least $1 - \zeta$ under the true data-generating distribution $\mathbb{P}_{\theta^*}^n$, the following holds for any $q_\theta \in \Gamma_\theta$ and $q_{\gamma(X^n)}(Z^n) \in \Gamma_{\gamma(X^n)}$:

$$\int D_\alpha^{(n)}(\theta, \theta^*)\, \hat{q}_{\theta,\alpha}(d\theta) \le \frac{\alpha}{n(1-\alpha)} \Psi_{n,\alpha}(q_\theta, q_{\gamma(X^n)}(Z^n)) + \frac{1}{n(1-\alpha)} \log\left(\frac{1}{\zeta}\right), \tag{24}$$

where

$$\Psi_{n,\alpha}(q_\theta, q_{\gamma(X^n)}(Z^n)) := \int_\Theta \left[\ell_n(\theta^*) - \widehat{\ell}_n^{\gamma}(\theta) \right] q_\theta(d\theta) + \frac{1}{\alpha} D(q_\theta \,\|\, p_\theta), \tag{25}$$

and

$$\widehat{\ell}_n^{\gamma}(\theta) := \int q_{\gamma(X^n)}(Z^n) \log\left(\frac{p(X^n \mid Z^n, \mu) \cdot p(Z^n \mid \lambda)}{q_{\gamma(X^n)}(Z^n)} \right) dZ^n. \tag{26}$$

To simplify the expression $\ell_n(\theta^*) - \widehat{\ell}_n^{\gamma}(\theta)$ in Equation (25), we proceed as follows. First note that by Bayes' rule:

$$p(Z^n \mid X^n, \theta^*) = \frac{p(X^n \mid Z^n, \mu^*) p(Z^n \mid \lambda^*)}{p(X^n \mid \theta^*)},$$

so that

$$\log p(Z^n \mid X^n, \theta^*) = \log p(X^n \mid Z^n, \mu^*) + \log p(Z^n \mid \lambda^*) - \log p(X^n \mid \theta^*). \tag{27}$$

Using this identity, we expand the KL divergence as

$$\begin{aligned}
&D(q_{\gamma(X^n)}(Z^n) \,\|\, p(Z^n \mid X^n, \theta^*)) \\
&= \int q_{\gamma(X^n)}(Z^n) \log\left(\frac{q_{\gamma(X^n)}(Z^n)}{p(Z^n \mid X^n, \theta^*)} \right) dZ^n \\
&= \int q_{\gamma(X^n)}(Z^n) \left[\log q_{\gamma(X^n)}(Z^n) - \log p(X^n \mid Z^n, \mu^*) - \log p(Z^n \mid \lambda^*) + \log p(X^n \mid \theta^*) \right] dZ^n \\
&= -\int q_{\gamma(X^n)}(Z^n) \log\left(\frac{p(X^n \mid Z^n, \mu^*) p(Z^n \mid \lambda^*)}{q_{\gamma(X^n)}(Z^n)} \right) dZ^n + \log p(X^n \mid \theta^*).
\end{aligned}$$

Using this, the definition of $\ell_n(\theta^*) = \log p(X^n \mid \theta^*)$, Equation (23), and Equation (26), we obtain:

$$\begin{aligned}
\ell_n(\theta^*) - \widehat{\ell}_n^{\gamma}(\theta) &= D(q_{\gamma(X^n)}(Z^n) \,\|\, p(Z^n \mid X^n, \theta^*)) \\
&\quad + \int q_{\gamma(X^n)}(Z^n) \log\left(\frac{p(X^n \mid Z^n, \mu^*) p(Z^n \mid \lambda^*)}{p(X^n \mid Z^n, \mu) p(Z^n \mid \lambda)} \right) dZ^n. \\
&\le n \Delta_{\textit{ASVIGap}}^2 + \int q_{\gamma(X^n)}(Z^n) \log\left(\frac{p(X^n \mid Z^n, \mu^*) p(Z^n \mid \lambda^*)}{p(X^n \mid Z^n, \mu) p(Z^n \mid \lambda)} \right) dZ^n.
\end{aligned} \tag{28}$$

Substituting Equation (28) into Equation (25), and then into Equation (24), completes the first step of the proof.

We now analyze the term $\int q_{\gamma(X^n)}(Z^n) \log\left(\frac{p(X^n|Z^n,\mu^*)p(Z^n|\lambda^*)}{p(X^n|Z^n,\mu)p(Z^n|\lambda)}\right) dZ^n$. *Specifically, we now control the expectation of it over* q_θ *in Equation (25). To that end, we choose the variational distribution* q_θ *as a renormalized restriction of the prior* $p_\theta = p_\lambda \cdot p_\mu$ *onto the high-probability events defined in the theorem statement. Define,*

$$\widetilde{Q}_\theta(\cdot) := \frac{P_\lambda\left[\cdot \cap \mathcal{B}_n^{\mathrm{VI}}(\lambda^*, \varepsilon_\lambda)\right] \cdot P_\mu\left[\cdot \cap \mathcal{B}_n^{\mathrm{ASVI}}(\mu^*, \varepsilon_\mu)\right]}{P_\lambda\left[\mathcal{B}_n^{\mathrm{VI}}(\lambda^*, \varepsilon_\lambda)\right] \cdot P_\mu\left[\mathcal{B}_n^{\mathrm{ASVI}}(\mu^*, \varepsilon_\mu)\right]}, \tag{29}$$

and let $\tilde{q}_\theta$ *denote its density with respect to* p_θ, *which satisfies* $\tilde{q}_\theta \ll p_\theta$.

By substituting $q_\theta = \tilde{q}_\theta$ *in Equation (25), we need to control the term*

$$T_\gamma(X^n) = \int_\Theta \left[\int q_{\gamma(X^n)}(Z^n) \log\left(\frac{p(X^n \mid Z^n, \mu^*)p(Z^n \mid \lambda^*)}{p(X^n \mid Z^n, \mu)p(Z^n \mid \lambda)}\right) dZ^n\right] q_\theta(d\theta). \tag{30}$$

Our goal is to show that $T_\gamma(X^n)$ *is bounded with high probability. Since this is a real-valued random variable measurable with respect to* $X^n \sim \mathbb{P}^n_{\theta^*}$, *we invoke Chebyshev's inequality.*

To apply this inequality effectively, we first derive explicit upper bounds on $\mathbb{E}_{\theta^*}[T_\gamma(X^n)]$ *and* $\mathrm{Var}_{\theta^*}(T_\gamma(X^n))$ *under the true data-generating distribution* $\mathbb{P}^n_{\theta^*}$. *To bound these moments for the integral in Equation (30), we compared* T_γ *with*

$$T(X^n) = \int_\Theta \left[\int \tilde{q}_{Z^n}(Z^n) \log\left(\frac{p(X^n \mid Z^n, \mu^*)p(Z^n \mid \lambda^*)}{p(X^n \mid Z^n, \mu)p(Z^n \mid \lambda)}\right) dZ^n\right] q_\theta(d\theta). \tag{31}$$

where the non amortized distribution $\tilde{q}_{Z^n}$ *is defined as,*

$$\tilde{q}_{Z^n}(Z^n) \propto P(X^n \mid \mu, Z^n)\, P(Z^n \mid \theta^*). \tag{32}$$

Observe that

$$\mathbb{E}_{\theta^*}(T_\gamma(X^n)) = \mathbb{E}_{\theta^*}(T_\gamma(X^n) - T(X^n)) + \mathbb{E}_{\theta^*}(T(X^n)).$$

and,

$$\mathrm{Var}_{\theta^*}(T_\gamma(X^n)) \le 2\,\mathrm{Var}_{\theta^*}(T_\gamma(X^n) - T(X^n)) + 2\,\mathrm{Var}_{\theta^*}(T(X^n)).$$

Then, by Lemma 4.1,

$$\mathbb{E}_{\theta^*}(T_\gamma(X^n) - T(X^n)) \le 4\left(n\Delta^2_{\mathit{ASVIGap}} + nf_\mu(n)\varepsilon_\mu^2 + f_\lambda(n)\varepsilon_\lambda^2\right) \tag{33}$$

and similarly,

$$\mathrm{Var}_{\theta^*}(T_\gamma(X^n) - T(X^n)) \le 8\left(n\Delta^2_{\mathit{ASVIGap}} + nf_\mu(n)\varepsilon_\mu^2 + f_\lambda(n)\varepsilon_\lambda^2\right). \tag{34}$$

Moreover, using Fubini's theorem, the expectation under $\mathbb{P}^n_{\theta^*}$ *becomes*

$$\mathbb{E}_{\theta^*}[T(X^n)] = \int_\Theta \mathbb{E}_{\theta^*}\left[\int q_{Z^n}(Z^n) \log\left(\frac{p(X^n \mid Z^n, \mu^*)\, p(Z^n \mid \lambda^*)}{p(X^n \mid Z^n, \mu)\, p(Z^n \mid \lambda)}\right) dZ^n\right] \tilde{q}_\theta(d\theta).$$

And since for all $\theta = (\mu, \lambda) \in \mathcal{B}_n^{\mathrm{VI}}$, *the inequalities in Equations (21) and (22) are satisfied, we obtain the bound,*

$$\mathbb{E}_{\theta^*}[T(X^n)] \le f_\lambda(n)\varepsilon_\lambda^2 + nf_\mu(n)\varepsilon_\mu^2. \tag{35}$$

We also require an upper bound on the variance of $T(X^n)$. *Similarly, we have that for all* $\theta = (\mu, \lambda) \in \mathcal{B}_n^{\mathrm{ASVI}}$, *using the Equations (21) and (22) the variance can be bounded as*

$$\mathrm{Var}_{\theta^*}(T(X^n)) \le 2\left(f_\lambda(n)\varepsilon_\lambda^2 + nf_\mu(n)\varepsilon_\mu^2\right). \tag{36}$$

Define,

$$\sigma_n^2 := 5(n\Delta^2_{\mathit{ASVIGap}} + f_\lambda(n)\varepsilon_\lambda^2 + nf_\mu(n)\varepsilon_\mu^2).$$

Now apply Chebyshev's inequality to deviation $D\sigma_n^2 - \mathbb{E}[T_\gamma(X^n)] \geq (D-1)\sigma_n^2$. The standard Chebyshev bound gives

$$\mathbb{P}_{\theta^*}^n \left(T_\gamma(X^n) > D\sigma_n^2 \right) \leq \frac{\mathrm{Var}(T_\gamma(X^n))}{(D-1)^2\sigma_n^4} \leq \frac{2\sigma_n^2}{(D-1)^2\sigma_n^4} = \frac{2}{(D-1)^2\sigma_n^2}.$$

Therefore,

$$\begin{aligned} &\mathbb{P}_{\theta^*}^n \left(T_\gamma(X^n) > D \cdot 5(n\Delta_{\textit{ASVIGap}}^2 + f_\lambda(n)\varepsilon_\lambda^2 + n f_\mu(n)\varepsilon_\mu^2) \right) \\ \leq &\frac{4}{25(D-1)^2 \left(f_\lambda(n)\varepsilon_\lambda^2 + n f_\mu(n)\varepsilon_\mu^2 \right)}. \end{aligned} \tag{37}$$

This establishes the desired high-probability control of $T(X^n)$. It reminds to observe that

$$D(\tilde{q}_\theta \,\|\, p_\theta) = -\log P_\lambda \left[\mathcal{B}_n^{\mathrm{VI}}(\lambda^*, \varepsilon_\lambda) \right] - \log P_\mu \left[\mathcal{B}_n^{\mathrm{VI}}(\mu^*, \varepsilon_\mu) \right]. \tag{38}$$

Substituting both Equation (37) and (38) into Equation (25) and then into Equation (24) yields the final bound in Theorem 1.

Lemma 4.1 *Suppose the variational approximation $q_{\gamma(X^n)} \in \mathcal{Q}_{\mathrm{AS}}$ satisfies the following property*

$$\inf_{q \in \mathcal{Q}_{\mathrm{AS}}} \frac{1}{n} D\left(q \,\|\, p(Z^n \mid X^n, \theta^*)\right) \leq \Delta_{\textit{ASVIGap}}^2. \tag{39}$$

Define the reference density

$$\tilde{q}_{X^n}(Z^n) \propto p(X^n \mid Z^n, \mu) \cdot p(Z^n \mid \lambda^*).$$

Let q_θ be a probability measure over Θ, and define the deviation term

$$\phi(\theta) := \int \left(q_{\gamma(X^n)}(Z^n) - \tilde{q}_{X^n}(Z^n) \right) \log \left(\frac{p(X^n \mid Z^n, \mu^*) \cdot p(Z^n \mid \lambda^*)}{p(X^n \mid Z^n, \mu) \cdot p(Z^n \mid \lambda)} \right) dZ^n.$$

Assume q_θ is supported on the set

$$\mathcal{B}_n := \mathcal{B}_n^{\mathrm{VI}}(\lambda^*, \varepsilon_\lambda) \cap \mathcal{B}_n^{\mathrm{VI}}(\mu^*, \varepsilon_\mu),$$

where for all $\theta = (\mu, \lambda) \in \mathrm{supp}(q_\theta)$:

$$\begin{gathered} D(p(Z^n \mid \lambda^*) \,\|\, p(Z^n \mid \lambda)) \leq f_\lambda(n)\varepsilon_\lambda^2,\ V(p(Z^n \mid \lambda^*) \,\|\, p(Z^n \mid \lambda)) \leq f_\lambda(n)\varepsilon_\lambda^2, \\ \max_{1 \leq i \leq n} \mathbb{E}_{Z^n \sim p(Z^n \mid X^n, \theta^*)} \left[D_i(\mu^*, \mu) \right] \leq f_\mu(n)\varepsilon_\mu^2,\ \max_{1 \leq i \leq n} \mathbb{E}_{Z^n \sim p(Z^n \mid X^n, \theta^*)} \left[V_i(\mu^*, \mu) \right] \leq f_\mu(n)\varepsilon_\mu^2. \end{gathered}$$

Then,

$$\left| \mathbb{E}_{\theta^*} \int_\Theta \phi(\theta)\, q_\theta(d\theta) \right| \leq 4 \left(n\Delta_{\textit{ASVIGap}}^2 + n f_\mu(n)\varepsilon_\mu^2 + f_\lambda(n)\varepsilon_\lambda^2 \right).$$

Proof 2 *We begin by expressing the deviation term of interest as*

$$\int_\Theta \phi(\theta)\, q_\theta(d\theta) = \int_\Theta \Big(\int \left(q_{\gamma(X^n)}(Z^n) - \tilde{q}_{X^n}(Z^n) \right) f_\theta(Z^n)\, dZ^n \Big) q_\theta(d\theta),$$

where $f_\theta(Z^n) = \log \left(\frac{p(X^n \mid Z^n, \mu^*) \cdot p(Z^n \mid \lambda^*)}{p(X^n \mid Z^n, \mu) \cdot p(Z^n \mid \lambda)} \right)$, and $\tilde{q}_{X^n}(Z^n) \propto p(X^n \mid Z^n, \mu) \cdot p(Z^n \mid \lambda^*)$.

Then, observe that by Fubini's Theorem

$$\int_\Theta \mathbb{E}_{\theta^*} \phi(\theta)\, q_\theta(d\theta) = \int_\Theta \mathbb{E}_{\theta^*} \Big(\int \left(q_{\gamma(X^n)}(Z^n) - \tilde{q}_{X^n}(Z^n) \right) f_\theta(Z^n)\, dZ^n \Big) q_\theta(d\theta).$$

For each fixed $\theta \in \Theta$, we now apply the classical Holder's inequality inequality,

$$\left| \int (q - \tilde{q}) f \right| \leq \left(\int \left(||f||_\infty \right)^2 \tilde{q} \right) \left(TV(q, \tilde{q}) \right).$$

Applying this to our integrand at each $\theta \in \Theta$, we obtain:

$$|\phi(\theta)| = \left| \int \left(q_{\gamma(X^n)}(Z^n) - \tilde{q}_{X^n}(Z^n) \right) f_\theta(Z^n)\, dZ^n \right| \leq \|f_\theta\|_\infty \cdot \left\| q_{\gamma(X^n)} - \tilde{q}_{X^n} \right\|_{\mathrm{TV}}.$$

This leads to,

$$\mathbb{E}_{\theta^*}|\phi(\theta)| \leq \mathbb{E}_{\theta^*}\left[\|f_\theta\|_{\infty^*} \cdot \left\| q_{\gamma(X^n)} - \tilde{q}_{X^n} \right\|_{\mathrm{TV}} \right].$$

Using Cauchy-Scwarz with respect to $\mathbb{E}_{\theta^*}$,

$$\mathbb{E}_{\theta^*}|\phi(\theta)| \leq \left(\mathbb{E}_{\theta^*} \|f_\theta\|_\infty^2 \right)^{1/2} \cdot \left(\mathbb{E}_{\theta^*} \left\| q_{\gamma(X^n)} - \tilde{q}_{X^n} \right\|_{\mathrm{TV}}^2 \right)^{1/2}.$$

Then using Pinsker's inequality

$$\left\| q_{\gamma(X^n)} - \tilde{q}_{X^n} \right\|_{\mathrm{TV}}^2 \leq \frac{1}{2} D\left(q_{\gamma(X^n)} \| \tilde{q}_{X^n} \right).$$

Now we apply Jensen's inequality to the product of the square roots,

$$\left| \int_\Theta \mathbb{E}_{\theta^*} \phi(\theta) q_\theta(d\theta) \right| \leq \sqrt{\int_\Theta \frac{1}{2} \mathbb{E}_{\theta^*} D\left(q_{\gamma(X^n)} \| \tilde{q}_{X^n} \right) q_\theta(d\theta)} \cdot \int_\Theta \left(\mathbb{E}_{\theta^*} \|f_\theta\|_\infty^2 \right)^{1/2} q_\theta(d\theta).$$

Then, for all $\theta \in \mathcal{B}_n$, we observe that the following inequalities hold,

$$\mathbb{E}_{\theta^*} \|f_\theta\|_\infty^2\, q_\theta(d\theta). \leq 2\left(n f_\mu(n) \varepsilon_\mu^2 + f_\lambda(n) \varepsilon_\lambda^2 \right),$$

Now, let $\pi^*(Z^n) := p(Z^n \mid X^n, \theta^*)$ and define the log-density ratio:

$$h(Z^n; \theta) := \log\left(\frac{p(X^n \mid Z^n, \mu^*) p(Z^n \mid \lambda^*)}{p(X^n \mid Z^n, \mu) p(Z^n \mid \lambda^*)} \right) = \log\left(\frac{p(X^n \mid Z^n, \mu^*)}{p(X^n \mid Z^n, \mu)} \right).$$

By construction, for fixed θ, we define

$$\tilde{q}_{X^n}(Z^n) := \frac{p(X^n \mid Z^n, \mu) \cdot p(Z^n \mid \lambda^*)}{Z_\theta}, \quad \textit{where } Z_\theta \textit{ normalizes } \tilde{q}_{X^n}.$$

We then write

$$D(q_{\gamma(X^n)} \,\|\, \tilde{q}_{X^n}) = \int q_{\gamma(X^n)}(Z^n) \log\left(\frac{q_{\gamma(X^n)}(Z^n)}{\tilde{q}_{X^n}(Z^n)} \right) dZ^n.$$

Now expand the denominator:

$$\log\left(\frac{q_{\gamma(X^n)}(Z^n)}{\tilde{q}_{X^n}(Z^n)} \right) = \log\left(\frac{q_{\gamma(X^n)}(Z^n)}{\pi^*(Z^n)} \cdot \frac{\pi^*(Z^n)}{\tilde{q}_{X^n}(Z^n)} \right) = \log\left(\frac{q_{\gamma(X^n)}(Z^n)}{\pi^*(Z^n)} \right) + h(Z^n; \theta) + \log Z_\theta.$$

Therefore,

$$D(q_{\gamma(X^n)} \,\|\, \tilde{q}_{X^n}) = D(q_{\gamma(X^n)} \,\|\, \pi^*) + \mathbb{E}_{q_{\gamma(X^n)}}[h(Z^n; \theta)] + \log Z_\theta.$$

Now integrate over q_θ,

$$\int D(q_{\gamma(X^n)} \,\|\, \tilde{q}_{X^n}) q_\theta(d\theta) = D(q_{\gamma(X^n)} \,\|\, \pi^*) + \int \mathbb{E}_{q_{\gamma(X^n)}}[h(Z^n; \theta)] q_\theta(d\theta) + \int \log Z_\theta\, q_\theta(d\theta).$$

Bound term 1: *By assumption,*

$$D(q_{\gamma(X^n)} \,\|\, \pi^*) \leq n \cdot \Delta_{\mathrm{ASVIGap}}^2.$$

Bound term 2: *We aim to bound*

$$\int \mathbb{E}_{q_{\gamma(X^n)}}[h(Z^n;\theta)]\, q_\theta(d\theta), \quad \textit{where } h(Z^n;\theta) := \log \frac{p(X^n \mid Z^n, \mu^*)}{p(X^n \mid Z^n, \mu)}.$$

We begin by adding and subtracting the expectation of $h(Z^n;\theta)$ under $\pi^*(Z^n) := p(Z^n \mid X^n, \theta^*)$:

$$\int \mathbb{E}_{q_{\gamma(X^n)}}[h(Z^n;\theta)]\, q_\theta(d\theta) = \int \mathbb{E}_{\pi^*}[h(Z^n;\theta)]\, q_\theta(d\theta) + \int \left(\mathbb{E}_{q_{\gamma(X^n)}}[h(Z^n;\theta)] - \mathbb{E}_{\pi^*}[h(Z^n;\theta)]\right) q_\theta(d\theta).$$

Step 1. *By assumption, for all $\theta \in \operatorname{supp}(q_\theta)$,*

$$\max_{1\le i\le n} \mathbb{E}_{Z^n\sim\pi^*}\left[D_i(\mu^*,\mu)\right] \le f_\mu(n)\varepsilon_\mu^2.$$

Therefore,

$$\int \mathbb{E}_{\pi^*}[h(Z^n;\theta)]\, q_\theta(d\theta) = \sum_{i=1}^n \int \mathbb{E}_{\pi^*}\left[\log \frac{p(X_i \mid Z_i, \mu^*)}{p(X_i \mid Z_i, \mu)}\right] q_\theta(d\theta) \le n f_\mu(n)\varepsilon_\mu^2.$$

Step 2. *Define the deviation*

$$\delta(\theta) := \mathbb{E}_{q_{\gamma(X^n)}}[h(Z^n;\theta)] - \mathbb{E}_{\pi^*}[h(Z^n;\theta)].$$

Apply the classical Holder's inequality

$$|\delta(\theta)| \le ||h(Z^n;\theta)||_\infty \cdot ||q_{\gamma(X^n)} - \pi^*||_{TV}.$$

Therefore

$$\left|\int_\Theta \mathbb{E}_{\theta^*}\delta(\theta) q_\theta(d\theta)\right| \le \sqrt{\int_\Theta \frac{1}{2}\mathbb{E}_{\theta^*} D\left(q_{\gamma(X^n)} \| \pi^*\right) q_\theta(d\theta)} \cdot \int_\Theta \left(\mathbb{E}_{\theta^*} \|h(\cdot,\theta)\|_\infty^2\right)^{1/2} q_\theta(d\theta).$$

As before, we get

$$\int_\Theta \left(\mathbb{E}_{\theta^*} \|h(\cdot,\theta)\|_\infty^2\right)^{1/2} q_\theta(d\theta) \le n f_\mu(n)\varepsilon_\mu^2.$$

Therefore,

$$\left|\int_\Theta \mathbb{E}_{\theta^*}\delta(\theta) q_\theta(d\theta)\right| \le \sqrt{2n\Delta^2_{\textit{ASVIGap}}} \cdot \sqrt{n f_\mu(n)\varepsilon_\mu^2} \le n f_\mu(n)\varepsilon_\mu^2 + n\Delta^2_{\textit{ASVIGap}}.$$

Combining both steps,

$$\int \mathbb{E}_{q_{\gamma(X^n)}}[h(Z^n;\theta)]\, q_\theta(d\theta) \le 3 \cdot \left(n f_\mu(n)\varepsilon_\mu^2 + n\Delta^2_{\textit{ASVIGap}}\right),$$

for a universal constant $C > 0$.

Bound term 3: *From the definition*

$$Z_\theta := \int p\left(X^n \mid Z^n, \mu\right) \cdot p\left(Z^n \mid \lambda^*\right) dZ^n = p\left(X^n \mid \mu, \lambda^*\right)$$

so

$$\log Z_\theta = \log p\left(X^n \mid \mu, \lambda^*\right)$$

Now, we compare this to the true marginal likelihood $p\left(X^n \mid \theta^*\right)$, and use the reverse KL inequality (also known as the log-integral inequality)

$$\log p\left(X^n \mid \theta^*\right) - \log p\left(X^n \mid \mu, \lambda^*\right) \le D\left(p\left(Z^n \mid \theta^*\right) \| p\left(Z^n \mid \mu, \lambda^*\right)\right)$$

Then, again by assumptions on $\theta \in \mathcal{B}_n$, this is bounded by

$$D\left(p\left(Z^n \mid \lambda^*\right) \| p\left(Z^n \mid \lambda\right)\right) + \sum_{i=1}^{n} \mathbb{E}_{\pi^*} D_i\left(\mu^*, \mu\right) \leq f_\lambda(n)\varepsilon_\lambda^2 + n f_\mu(n)\varepsilon_\mu^2$$

So

$$\log p\left(X^n \mid \mu, \lambda^*\right) \leq \log p\left(X^n \mid \theta^*\right) + f_\lambda(n)\varepsilon_\lambda^2 + n f_\mu(n)\varepsilon_\mu^2$$

and integrating over q_θ, we get,

$$\int \log Z_\theta q_\theta(d\theta) \leq \log p\left(X^n \mid \theta^*\right) + f_\lambda(n)\varepsilon_\lambda^2 + n f_\mu(n)\varepsilon_{\mu^*}^2$$

Final bound

$$\int_\Theta D(q_{\gamma(X^n)} \,\|\, \tilde{q}_{X^n}) q_\theta(d\theta) \leq 4\left(n\Delta_{\textit{ASVIGap}}^2 + n f_\mu(n)\varepsilon_\mu^2 + f_\lambda(n)\varepsilon_\lambda^2\right).$$

Thus,

$$\left|\int_\Theta \mathbb{E}_{\theta^*}\phi(\theta)\, q_\theta(d\theta)\right| \leq \left(8\left(n\Delta_{\textit{ASVIGap}}^2 + n f_\mu(n)\varepsilon_\mu^2 + f_\lambda(n)\varepsilon_\lambda^2\right)\right)^{1/2}\left(2\left(n f_\mu(n)\varepsilon_\mu^2 + f_\lambda(n)\varepsilon_\lambda^2\right)\right)^{1/2},$$

and the result is obtained.

5 Proof of Corollary 1

Proof 3 *We work under the setup of the truncated linear-Gaussian latent state-space model, where $Z_i \in \mathbb{R}^d$ evolves according to the dynamics*

$$Z_1 \sim \mathcal{TN}_{[-R_1,R_1]^d}(0, \tau^2 I_d), \quad Z_i = AZ_{i-1} + b + \epsilon_i, \quad \epsilon_i \sim \mathcal{TN}_{[-R_1,R_1]^d}(0, \tau^2 I_d).$$

Observations are given by

$$X_i = \omega + f(Z_i) + \eta_i, \quad \eta_i \sim \mathcal{TN}_{[-R_2,R_2]^d}(0, \sigma^2 I_d),$$

where $f : \mathbb{R}^d \to \mathbb{R}^d$ is a fixed nonlinear function.

Step 1: Events bounds. *We begin by proving that $f_\lambda(n) = O(n)$ and $f_\mu(n) = O(1)$, since Theorem 1 requires these scaling behaviors in order to bound the variational risk. To this end, remember that the quantities $f_\lambda(n)$ and $f_\mu(n)$ are defined in the concentration events*

$$\begin{aligned}
\mathcal{B}_n^{\mathrm{VI}}(\lambda^*, \varepsilon_\lambda) := \{\theta = (\mu, \lambda) : & D\big(p(Z^n \mid \lambda^*) \| p(Z^n \mid \lambda)\big) \leq f_\lambda(n)\varepsilon_\lambda^2, \\
& V\big(p(Z^n \mid \lambda^*) \| p(Z^n \mid \lambda)\big) \leq f_\lambda(n)\varepsilon_\lambda^2\}, \\
\mathcal{B}_n^{\mathrm{VI}}(\mu^*, \varepsilon_\mu) := \Big\{\theta = (\mu, \lambda) : & \max_{1\leq i\leq n} \mathbb{E}_{Z^n \sim p(Z^n \mid X^n, \theta^*)} D_i(\mu^*, \mu) \leq f_\mu(n)\varepsilon_\mu^2, \\
& \max_{1\leq i\leq n} \mathbb{E}_{Z^n \sim p(Z^n \mid X^n, \theta^*)} V_i(\mu^*, \mu) \leq f_\mu(n)\varepsilon_\mu^2\Big\}.
\end{aligned}$$

First, consider the KL divergence term $D(p(Z^n \mid \lambda^*) \,\|\, p(Z^n \mid \lambda))$. Using the chain rule for KL divergence applied to the latent Markov model, we have

$$D(p(Z^n \mid \lambda^*) \| p(Z^n \mid \lambda)) = \sum_{i=1}^{n} D(p(Z_i \mid Z_{i-1}, \lambda^*) \| p(Z_i \mid Z_{i-1}, \lambda)),$$

where each transition $p(Z_i \mid Z_{i-1}, \lambda^*)$ is a truncated Gaussian $\mathcal{TN}_{[-R_1,R_1]^d}(A^* Z_{i-1} + b^*, (\tau^*)^2 I_d)$ under $\lambda^* = (A, b)$, and $p(Z_i \mid Z_{i-1}, \lambda)$ is a truncated Gaussian $\mathcal{TN}_{[-R_1,R_1]^d}(AZ_{i-1} + b, \tau^2 I_d)$ under $\lambda = (A, b)$. Each term in the sum

is a KL divergence between two truncated Gaussian distributions over a bounded domain. Although there is no closed-form expression for the KL divergence between truncated Gaussians, upper bounds can be obtained via the standard KL divergence between the corresponding untruncated Gaussians, which gives

$$D(p(Z_i \mid Z_{i-1}, \lambda^*) \| p(Z_i \mid Z_{i-1}, \lambda)) \leq C \cdot \left(\|A^* - A\|_F^2 + \|b^* - b\|_2^2 + ((\tau^*)2 - \tau^2)^2 \right),$$

for some constant C depending on the truncation region. Now, under the concentration event $\lambda \in \mathcal{B}_n^{\mathrm{VI}}(\lambda^*, \varepsilon_\lambda)$, we have

$$\|A^* - A\|_F^2 + \|b^* - b\|_2^2 + ((\tau^*)2 - \tau^2)^2 \leq \varepsilon_\lambda^2,$$

so each KL term is $O(\varepsilon_\lambda^2)$ and same for the Variance term, and summing over n terms yields

$$f_\lambda(n) \cdot \varepsilon_\lambda^2 \asymp n \cdot \varepsilon_\lambda^2 \Rightarrow f_\lambda(n) = O(n).$$

We now analyze the term $f_\mu(n)$. Recall that for each $\theta = (\mu, \lambda)$, the function $f_\mu(n)\varepsilon_\mu^2$ bounds the average predictive discrepancy in the observation model

$$\max_{1 \leq i \leq n} \mathbb{E}_{Z^n \sim p(Z^n \mid X^n, \theta^*)} \left[D_i(\mu^*, \mu) \right] \leq f_\mu(n) \varepsilon_\mu^2.$$

Here, $D_i(\mu^*, \mu)$ represents the KL divergence between $p(X_i \mid Z_i, \mu^*)$ and $p(X_i \mid Z_i, \mu)$. Under our generative model, these are both truncated Gaussians

$$p(X_i \mid Z_i, \mu^*) = \mathcal{TN}_{[-R_2, R_2]^d}(\omega + f(Z_i), (\sigma^*)^2 I_d),$$

$$p(X_i \mid Z_i, \mu) = \mathcal{TN}_{[-R_2, R_2]^d}(\omega + f(Z_i), \sigma^2 I_d),$$

so the KL divergence between them is again controlled by the sum of the squared distances $\|\omega^* - \omega\|_2^2$ and $((\sigma^*)^2 - (\sigma)^2)^2$. Similarly for the Variance term, the same is obtained.

Under the concentration event $\mu \in \mathcal{B}_n^{\mathrm{VI}}(\mu^*, \varepsilon_\mu)$, this squared error is $O(\varepsilon_\mu^2)$, and the KL divergence and V variance are bounded independently of n, yielding

$$f_\mu(n) \cdot \varepsilon_\mu^2 = O(\varepsilon_\mu^2) \Rightarrow f_\mu(n) = O(1).$$

Step 2: Prior Mass Bound. *We now turn to bounding the prior mass term appearing in Theorem 1,*

$$\frac{1}{n} \left(\log P_\lambda \left(\mathcal{B}_n^{\mathrm{VI}}(\lambda^*, \varepsilon_\lambda) \right) + \log P_\mu \left(\mathcal{B}_n^{\mathrm{VI}}(\mu^*, \varepsilon_\mu) \right) \right).$$

Recall we set

$$\varepsilon_\lambda = \varepsilon_\mu = \frac{(\log n)^\beta}{n}, \quad \textit{for some } \beta > 0.$$

From the previous step, we have shown that $f_\lambda(n) = O(n)$ and $f_\mu(n) = O(1)$ hold under our model and the definition of the KL concentration events $\mathcal{B}_n^{\mathrm{VI}}(\cdot)$. Additionally, the same KL/variance bounds imply

$$D(p(Z^n \mid \lambda^*) \,\|\, p(Z^n \mid \lambda)) \leq C_1(R_1, R_2) f_\lambda(n) \|\lambda - \lambda^*\|^2,$$
$$\max_{1 \leq i \leq n} \mathbb{E}_{Z^n \sim p(Z^n \mid X^n, \theta^*)} [D_i(\mu^*, \mu)] \leq C_2(R_1, R_2) f_\mu(n) \|\mu - \mu^*\|^2,$$

for constants $C_1(R_1, R_2), C_2(R_1, R_2) > 0$. These imply that

$$\left\{ \|\lambda - \lambda^*\|^2 \leq \frac{\varepsilon_\lambda^2}{C_1 f_\lambda(n)} \right\} \subseteq \mathcal{B}_n^{\mathrm{VI}}(\lambda^*, \varepsilon_\lambda), \quad \left\{ \|\mu - \mu^*\|^2 \leq \frac{\varepsilon_\mu^2}{C_2 f_\mu(n)} \right\} \subseteq \mathcal{B}_n^{\mathrm{VI}}(\mu^*, \varepsilon_\mu).$$

Now assume the priors P_λ and P_μ admit continuous densities that are bounded away from zero in neighborhoods around λ^* and μ^*. Then standard volume bounds for Euclidean balls in $\mathbb{R}^{d_\lambda}$ and $\mathbb{R}^{d_\mu}$ yield

$$P_\lambda \left(\mathcal{B}_n^{\mathrm{VI}}(\lambda^*, \varepsilon_\lambda) \right) \gtrsim \left(\frac{\varepsilon_\lambda^2}{C_1 f_\lambda(n)} \right)^{d_\lambda/2},$$

$$P_\mu \left(\mathcal{B}_n^{\mathrm{VI}}(\mu^*, \varepsilon_\mu) \right) \gtrsim \left(\frac{\varepsilon_\mu^2}{C_2 f_\mu(n)} \right)^{d_\mu/2}.$$

Taking logarithms and dividing by n, we get

$$\frac{1}{n}\left(\log P_\lambda(\mathcal{B}_n^{\mathrm{VI}}) + \log P_\mu(\mathcal{B}_n^{\mathrm{VI}})\right) \geq -\frac{C'}{n}\left(d_\lambda \log\left(\frac{n^2 f_\lambda(n)}{(\log n)^{2\beta}}\right) + d_\mu \log\left(\frac{n^2 f_\mu(n)}{(\log n)^{2\beta}}\right)\right),$$

for some universal constant $C' > 0$ depending on R_1, R_2. Substituting $f_\lambda(n) = O(n)$, $f_\mu(n) = O(1)$, we conclude

$$\frac{1}{n}\left(\log P_\lambda(\mathcal{B}_n^{\mathrm{VI}}) + \log P_\mu(\mathcal{B}_n^{\mathrm{VI}})\right) \gtrsim -\frac{(\log n)^{\beta'}}{n},$$

for some constant $\beta' > 0$ depending on β, d_λ, and d_μ. This completes the second required component of Theorem 1, the $\Delta^2_{\mathit{VIGap}}$.

Step 3: Bounding the Amortization Gap $\Delta^2_{\mathit{ASVIGap}}$. *We now complete the proof of Corollary 1 by bounding the amortization gap term*

$$\Delta^2_{\mathit{ASVIGap}} := \inf_{q_{\gamma(X^n)} \in \mathcal{Q}_{AS}} \frac{1}{n} D\left(q_{\gamma(X^n)} \,\|\, p(Z^n \mid X^n, \theta^*)\right),$$

appearing in Theorem 1. This KL divergence measures how well the amortized variational approximation $q_{\gamma(X^n)}$ matches the latent posterior $p(Z^n \mid X^n, \theta^*)$.

We consider the structured amortized variational family,

$$q_{\gamma(X^n)}(Z^n) = q_{\gamma(X_1)}(Z_1) \cdot \prod_{i=2}^{n} q_{\gamma(X_i)}(Z_i \mid Z_{i-1}),$$

where each conditional is a truncated Gaussian distribution:

$$q_{\gamma(X_i)}(Z_i \mid Z_{i-1}) = \mathcal{TN}_{[-R_1,R_1]^d}\left(Z_i;\, A_\gamma(X_i)\, Z_{i-1} + b_\gamma(X_i),\, \Sigma_\gamma(X_i)\left(I_d + 2 \cdot \frac{Z_{i-1} Z_{i-1}^\top}{1 + \|Z_{i-1}\|_2^2}\right)\right).$$

The encoder $\gamma : \mathbb{R}^d \to \mathbb{R}^{d\times d} \times \mathbb{R}^d \times \mathbb{S}^d_{++}$ is implemented via a fully connected ReLU neural network of architecture $\gamma \in \mathcal{F}(L, r)$, with one of the two configurations,

- $L \asymp \log(n)$, $r \asymp n^{1/(2(2p+1))}$, *or*
- $L \asymp \log(n) \cdot n^{1/(2(2p+1))}$, $r = O(1)$,

for a sufficiently large $p > 0$. We denote by $m = d^2 + 2d$ the total number of outputs of the encoder, corresponding to the entries of $A_\gamma(X_i)$, $b_\gamma(X_i)$, and the diagonal of $\Sigma_\gamma(X_i)$.

We assume that the true transition parameters (A^*, b^*, Σ^*) are constant across time and belong to the approximation class $\mathcal{H}(1, \{(p, 1)\})$, meaning each entry can be approximated with accuracy δ using ReLU networks of width and depth depending on the smoothness parameter p.

By Theorem 1 of Kohler and Langer (2021), the ReLU network class $\mathcal{F}(L, r)$ achieves an L^2-approximation error rate of

$$\|\gamma(X_i) - \gamma^*(X_i)\|^2 = O\left(\log^c(n) \cdot n^{-2p/(2p+1)}\right),$$

uniformly for all $i = 1, \ldots, n$, where $\gamma^*(X_i)$ denotes the true parameter map (i.e., mapping X_i to (A^*, b^*, Σ^*)). This is a uniform rate over the compact support of $X_i \in [-R_2, R_2]^d$, which is valid due to the truncation in the observation model.

Since m total outputs need to be approximated across n time steps, and the variational approximation is entry-wise in the encoder, the total approximation error (per sample) satisfies

$$\frac{1}{n} D\left(q_{\gamma(X^n)} \,\|\, p(Z^n \mid X^n, \theta^*)\right) \lesssim m \cdot \log^c(n) \cdot n^{-2p/(2p+1)}.$$

Hence, the amortization gap satisfies

$$\Delta^2_{\mathit{ASVIGap}} \lesssim m \cdot \log^c(n) \cdot n^{-2p/(2p+1)},$$

as claimed.

References

Kohler, M. and Langer, S. (2021). On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics*, 49(4):2231–2249.

Srivastava, N., Mansimov, E., and Salakhudinov, R. (2015). Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852. PMLR.

Yang, Y., Pati, D., and Bhattacharya, A. (2020). α-variational inference with statistical guarantees. *The Annals of Statistics*, 48(2):886–905.