Amortized Structural Variational Inference

Anonymous Author
Anonymous Institution

Abstract

Variational inference (VI) is a popular method for approximate Bayesian inference but can scale poorly and requires re-optimization for new data. Amortized variational inference (AVI) addresses this by learning a global inference map, though standard mean-field AVI often suffers from large variational and amortization gaps due to independence assumptions. We propose amortized structural variational inference (ASVI), which incorporates structural dependencies among latent variables through neural architectures that encode local neighborhood structure. ASVI reduces both variational and amortization gaps while retaining scalability. Simulations and real data experiments show that ASVI improves predictive accuracy and posterior fidelity over AVI, matching the performance of structured VI at lower computational cost.

1 Introduction

Variational inference (VI) [1, 2] 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 [3]. 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) [4, 5, 6, 7], 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 [8, 9], finite mixture models [10, 11], state-space models [12, 13], 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 [14]. Recent efforts to reduce the amortization gap include the use of more expressive network architectures, regularization techniques, and hybrid optimization methods that combine amortized inference with per-instance fine-tuning (e.g., [15, 16, 17]). However, many of these methods remain heuristic or introduce significant additional computational cost.

A parallel line of work has explored the construction of

Preliminary work. Under review by AISTATS 2026. Do not distribute.

richer variational families to improve approximation fidelity, particularly in Bayesian models with dependent latent variables. For instance, [18] introduced structured variational approximations to better capture dependencies among latent variables. Similarly, [19] incorporated structural information into variational inference for latent state models, with a focus on the state-space model. [20] 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 [21] and temporal VAEs with recurrent architectures [22], 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, such as its depth and width, 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 $p(X^n,\, Z^n \,|\, \theta) = p(Z^n \,|\, \theta) \cdot \prod_{i=1}^n p(X_i \,|\, \theta, Z_i)$, where $p(X_i \,|\, \theta, Z_i)$ denotes the likelihood of the individual observation X_i given the corresponding latent variable Z_i and the global parameter θ, and $p(Z^n \,|\, \theta)$ 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 $p(\theta, Z^n \,|\, X^n)$ by selecting a distribution $\widehat{Q}$ from a chosen variational family $\mathcal{Q}$ that minimizes the Kullback–Leibler (KL) divergence to the true posterior:

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

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, [23]), 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 [24], 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., [25]) 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, [26]), 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 [19]. 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 $p(Z^n \,|\, \theta)$ [19], 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:

$$\begin{aligned} &X_i \mid Z_i \sim p_\mu(X_i \mid Z_i), \;\; i = 1, \ldots, n, \\ &Z^n = (Z_1, \ldots, Z_n) \sim p_\lambda(Z_1, \ldots, Z_n), \end{aligned}$$

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 $q(Z_i \,|\, Z_{C_i \setminus \{i\}})$ in the variational family $\mathcal{Q}_{\mathrm{AS}}$, using the corresponding observations xC_i.

To illustrate $\mathcal{Q}_{\mathrm{AS}}$, we focus on general state-space models [27, 28], specified by

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

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 [29, 19] 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 [30], 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 [19] 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.

Theorem 1 *Let $\widehat{q}_{\theta,\alpha}$ denote the amortized variational posterior obtained from the α-variational inference framework (see Appendix B) using the variational family $\mathcal{Q}_{AS}$, 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. 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*

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

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

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

where 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 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 $\mathcal{Q}_S = \{q : q(\theta, Z^n) = q_\theta(\theta) \cdot q_{Z^n}(Z^n)\}$.

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 [19], 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 $\Delta^2_{\text{ASVIGap}} \lesssim m \cdot \log^c(n)\, n^{-\frac{2p}{2p+1}}$. 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:*

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

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 [19]. 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 [19], 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 [19] 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. 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.

Specifically, we consider the following variational families: the mean-field VI $\mathcal{Q}_{\text{MF}} = \{q : q(\theta, Z^n) = q(\theta)\prod_{i=1}^n q_i(Z_i)\}$; the naive amortized VI $\mathcal{Q}_{\text{const}} = \{q : q(\theta, Z^n) = q_0(\theta)\prod_{i=1}^n q_{\text{const}}(Z_i)\}$; the standard amortized VI $\mathcal{Q}_{\text{A}} = \{q : q(\theta, Z^n) = q_0(\theta)\prod_{i=1}^n q_{\gamma(X_i)}(Z_i)\}$; the neighborhood-based amortized VI $\mathcal{Q}_{\text{AN}} = \{q : q(\theta, Z^n) = q_0(\theta)\prod_{i=1}^n q_{\gamma(X_{C_i})}(Z_i)\}$; the structured VI without amortization $\mathcal{Q}_{\text{S}} = \{q : q(\theta, Z^n) = q_0(\theta)q_1(Z_1)\prod_{i=2}^n q_i(Z_i \mid Z_{i-1})\}$; and the amortized structured VI $\mathcal{Q}_{\text{AS}} = \{q : q(\theta, Z^n) = q_0(\theta)q_{\gamma(X_{C_1})}(Z_1)\prod_{i=2}^n q_{\gamma(X_{C_i})}(Z_i \mid Z_{i-1})\}$, where each C_i denotes a neighborhood of X_i with $|C_i| > 1$.

Note that $\mathcal{Q}_{\text{const}} \subset \mathcal{Q}_{\text{A}} \subset \mathcal{Q}_{\text{AN}} \subset \mathcal{Q}_{\text{MF}}$ and $\mathcal{Q}_{\text{AS}} \subset \mathcal{Q}_{\text{S}}$. We specify $q(\cdot)$ to be Gaussian and model $\gamma(\cdot)$ using a two-layer ReLU feedforward network. For $\mathcal{Q}_{\text{A}}$ and $\mathcal{Q}_{\text{AN}}$, we set $q_{\gamma(\cdot)}(z) = \mathcal{N}\big(z; \mu_\gamma(\cdot), \Sigma_\gamma(\cdot)\big)$. For $q \in \mathcal{Q}_{\text{S}}$, one has $q(\theta, Z^n) = q_1(Z_1)\prod_{i=2}^n q_i(Z_i \mid Z_{i-1}) = \mathcal{N}(Z_1; b_1, \Sigma_1)\prod_{i=2}^n \mathcal{N}(Z_i; b_i + A_{i-1}Z_{i-1}, \Sigma_i)$. For the amortized structured family $\mathcal{Q}_{\text{AS}}$, the triplets (A_i, b_i, Σ_i), $i = 1, \ldots, n$, are amortized with $q(\theta, Z^n) = \mathcal{N}(Z_1; a_\gamma(X_1), \Sigma_\gamma(X_1))\prod_{i=2}^n \mathcal{N}(Z_i; b_\gamma(X_i) + A_\gamma(X_i)\cdot Z_{i-1}, \Sigma_\gamma(X_i))$. The algorithm is implemented with the fractional posterior using $\alpha = 0.99$.

4.1 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.

Algorithm 1 Amortized structured VI optimization

Input: Data points $X^n = \{X_1, X_2, \ldots, X_n\}$.
Output: Parameters $\{A_z, b_z, \Sigma_z, \mu_\theta, \Sigma_\theta\}$, Approximation $q(Z_i|Z_{i-1})q(\theta) = \mathcal{N}(b_z + A_z Z_{i-1}, \Sigma_z)\mathcal{N}(\mu_\theta, \Sigma_\theta)$ for $i \geq 2$ and $q(Z_1) = N(b_z, \Sigma_z)$.
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 $\{\theta_j\}_{j=1}^m$ from $\mathcal{N}(\mu_\theta, \Sigma_\theta)$.
  for i in 1, ..., n **do**
    Sample Z_1 from $q(Z_{1j}) = \mathcal{N}(a_z, \Sigma_z(X_1))$ and $Z_i = \{Z_{ij}\}_{j=1}^m$ from $\mathcal{N}(b_z + A_z Z_{i-1}, \Sigma_z(X_i))$ for $i \geq 2$.
    Compute $\mathcal{L}(X_i; \theta, Z_i) = \sum_{j=1}^m \big[\log p_\alpha(X_i, Z_{ij}, \theta_j) - \log q(\theta_j, Z_{ij}|X_i)\big]$.
    Compute $\nabla \log \mathcal{L}(\theta_i, Z_i)$ with respect to $\{A_z, b_z, \Sigma_z, \mu_\theta, \Sigma_\theta\}$.
  end for
  $\widehat{\nabla} \log \mathcal{L} = \frac{1}{n}\sum_i \nabla \log \mathcal{L}(\theta_i, Z_i)$.
  Update $\{A_z, b_z, \Sigma_z, \mu_\theta, \Sigma_\theta\}$ using $\widehat{\nabla}\mathcal{L}$.
end while
return $\{A_z, b_z, \Sigma_z, \mu_\theta, \Sigma_\theta\}$.

4.2 AR(1) model

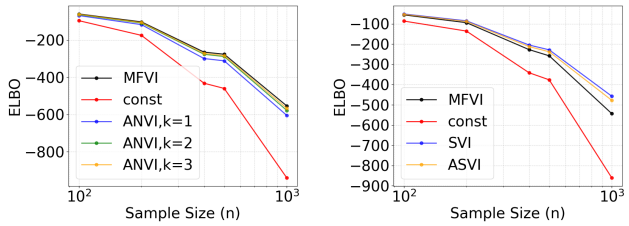
As our first example, we consider the simple AR(1) latent variable model and conduct a comprehensive simulation study to demonstrate that the ASVI framework is computationally efficient. Our results indicate that incorporating structural information improves estimation accuracy by reducing the amortization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

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:

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

Here we set $\theta = 2$, $\tau = 0.5$, 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.

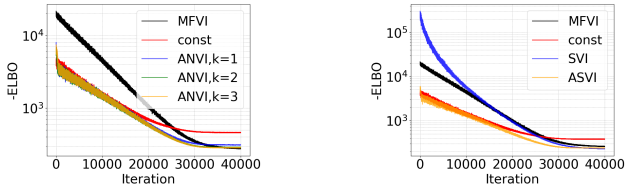
In Figure 1a, we focus on the mean-field variational family, with the goal of illustrating that incorporating structural



(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.

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

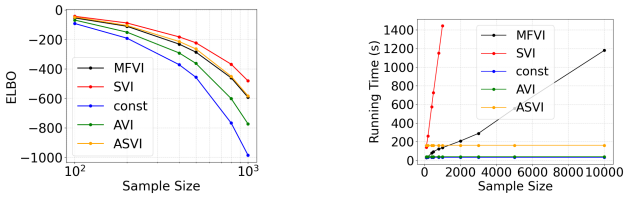
Figure 2: The optimization sample paths of different methods. A small value at the final iteration is preferred.

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 the MF variational family requires the largest number of optimization steps for the $-$ELBO to converge (as indicated by the plateau of the curves). In contrast, the ANVI methods converge much more quickly, highlighting the computational gains achieved through amortization. Figure 2b presents a similar pattern: the optimization trajectories show that ASVI converges substantially faster than both MFVI and SVI.

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.



(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

4.3 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:

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

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

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 $c_k \equiv 0$.

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 [31]. 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) $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|X_i - \widehat{X}_i\|_2^2}$, where $\widehat{X}_i$ is the prediction by the posterior mean of Z_i. Results from Table 1 show that increasing the amortization window k steadily improves both fit and prediction: ELBO increases, while MSE (global) and MSE(pred) decrease, bringing ASVI performance close to that of SVI. FlowASVI underperforms ASVI on all three metrics, and MFVI remains the weakest overall. The gains from enlarging k diminish beyond $k = 2$, with $k = 3$ providing the best ASVI performance, though still slightly below SVI in terms of ELBO.

Table 1: AR(p) performance summary.

Method	ELBO	MSE (global)	MSE (pred)
MFVI	-873.48	0.2254	0.572
SVI	**-811.25**	**0.0082**	**0.561**
ASVI_k1	-833.40	0.0298	0.597
ASVI_k2	-835.72	0.0267	0.582
ASVI_k3	-816.71	0.0314	0.570
FlowASVI_k1	-838.07	0.0948	0.602
FlowASVI_k2	-835.95	0.0996	0.597
FlowASVI_k3	-835.40	0.1048	0.581

4.4 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 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 2: Nonlinear AR(2) performance summary.

Method	ELBO	MSE (global)	MSE (pred)
MFVI	-1153.209	1.68690	0.1208
SVI	-215.947	0.18612	0.0684
ASVI	-205.022	**0.02552**	**0.0516**
FlowASVI	**-201.306**	0.08148	0.1577

Across Table 2, 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.

4.5 Real data example

In the real data analysis, 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 3, ASVI achieves the lowest test MSE among all priors, demonstrating superior one-step forecasting without access to the current image.

Prior	Train MSE	Test MSE
Random	0.05401	0.05452
ASVI	0.01495	**0.01761**
Gaussian	0.01813	0.01979
Flow	**0.00600**	0.02062

Table 3: One-step prediction MSE comparison.

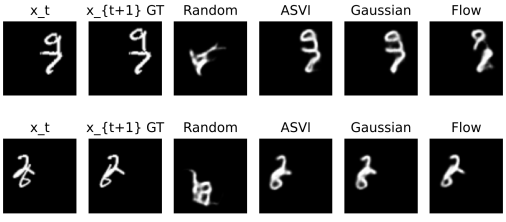


Figure 4: Bottom row: example of predicted moving digit from its previous frame using different methods on the training dataset. Top row: corresponding results on the test dataset.

References

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

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

[3] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv e-prints*, page arXiv:1312.6114, December 2013.

[4] Ankush Ganguly, Sanjana Jain, and Ukrit Watchareeruetai. Amortized variational inference: A systematic review. *Journal of Artificial Intelligence Research*, 78:167–215, 2023.

[5] Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.

[6] Charles C Margossian and David M Blei. Amortized variational inference: when and why? *arXiv preprint arXiv:2307.11018*, 2023.

[7] Abhinav Agrawal and Justin Domke. Amortized variational inference for simple hierarchical models. *Advances in Neural Information Processing Systems*, 34:21388–21399, 2021.

[8] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[9] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia tools and applications*, 78:15169–15211, 2019.

[10] Stephen J Roberts, Dirk Husmeier, Iead Rezek, and William Penny. Bayesian approaches to gaussian mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1133–1142, 1998.

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

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

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

[14] Chris Cremer, Xuechen Li, and David Duvenaud. Inference suboptimality in variational autoencoders. In *International conference on machine learning*, pages 1078–1086. PMLR, 2018.

[15] Mingtian Zhang, Peter Hayes, and David Barber. Generalization gap in amortized inference. *Advances in neural information processing systems*, 35:26777–26790, 2022.

[16] Minyoung Kim and Vladimir Pavlovic. Reducing the amortization gap in variational autoencoders: A bayesian random function approach. *arXiv preprint arXiv:2102.03151*, 2021.

[17] Rahul Krishnan, Dawen Liang, and Matthew Hoffman. On the challenges of learning with inference networks on sparse, high-dimensional data. In Amos Storkey and Fernando Perez-Cruz, 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, 09–11 Apr 2018.

[18] Eric P Xing, Michael I Jordan, and Stuart Russell. A generalized mean field algorithm for variational inference in exponential families. *arXiv preprint arXiv:1212.2512*, 2012.

[19] Honggang Wang, Anirban Bhattacharya, Debdeep Pati, and Yun Yang. Structured variational inference in bayesian state-space models. In *International Conference on Artificial Intelligence and Statistics*, pages 8884–8905. PMLR, 2022.

[20] Yixiu Zhao and Scott Linderman. Revisiting structured variational autoencoders. In *International Conference on Machine Learning*, pages 42046–42057. PMLR, 2023.

[21] Thomas Kipf and Max Welling. Variational graph auto-encoders. *NeurIPS Workshop on Bayesian Deep Learning (NeurIPS BDL)*, abs/1611.07308, 2016.

[22] Marco Fraccaro, Søren Kaae Sønderby, Ulrich Paquet, and Ole Winther. 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, 2016. Curran Associates Inc.

[23] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[24] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *the Journal of machine Learning research*, 14(1):1303–1347, 2013.

[25] Charles C Margossian and David M Blei. Amortized variational inference: When and why? In *Uncertainty in Artificial Intelligence*, pages 2434–2449. PMLR, 2024.

[26] Jakub Tomczak and Max Welling. Vae with a vampprior. In *International conference on artificial intelligence and statistics*, pages 1214–1223. PMLR, 2018.

[27] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[28] Yong Zeng. *State-Space Models applications in economics and finance*. Springer, 2013.

[29] Yun Yang, Debdeep Pati, and Anirban Bhattacharya. α-variational inference with statistical guarantees. *The Annals of Statistics*, 48(2):886–905, 2020.

[30] Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[31] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

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) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [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]