Amortized Variational Inference: When and Why? (Supplemental Material)

A MISSING PROOFS

We provide proofs for all statements in §3.

A.1 CAVI RULE

Lemma 3.1 follows from the coordinate-ascent VI update rule for F-VI [Blei et al., 2017, Eq. 17], which tells us how to choose $q(z_n\,;\,\nu_n)$ to minimize the KL-divergence, while maintaining the other factors in the approximating distribution fixed. Specifically, suppose ν_0 and $\boldsymbol{\nu}_{-n}$ are fixed. Then the optimal variational parameter $\nu_n^\star$ for n^{th} factor verifies

$$q(z_n\,;\,\nu_n^\star) \propto \exp\left\{\mathbb{E}_{q(\theta\,;\,\nu_0)}\left[\mathbb{E}_{q(\mathbf{z}_{-n}\,;\,\nu)}\left[\log p(\theta, \mathbf{z}, \mathbf{x})\right]\right]\right\}. \tag{1}$$

We now apply this rule to the optimal solution, i.e. we set $\nu_0 = \nu_0^*$ and $\boldsymbol{\nu}_{-n} = \boldsymbol{\nu}_{-n}^*$. Then, minimizing the KL-divergence, $\nu_n^\star = \nu_n^*$ and the desired result follows. □

A.2 EXISTENCE OF AN IDEAL INFERENCE FUNCTION AND SIMPLE HIERARCHICAL MODELS

Theorem 3.3 states that the existence of an ideal inference function for a standard latent variable model (Definition 3.2) is, in general, equivalent to $p(\theta, \mathbf{z}, \mathbf{x})$ being a simple hierarchical model (Eq. 1).

We first prove item (1). Suppose $p(\theta, \mathbf{z}, \mathbf{x})$ is a simple hierarchical model. Applying the CAVI rule (Lemma 3.1) to Eq. 1,

$$\begin{aligned} q(z_n\,;\,\nu^*) &\propto \exp\left\{\mathbb{E}_{q(\theta\,;\,\nu_0^*)}\left[\mathbb{E}_{q(\mathbf{z}_{-n}\,;\,\nu^*)}\left[\log p(\theta) + \sum_{j=1}^{n} \log p(z_j \mid \theta) + \log p(x_j \mid z_j, \theta)\right]\right]\right\} \\ &\propto \exp\left\{\mathbb{E}_{q(\theta\,;\,\nu_0^*)}\left[\mathbb{E}_{q(\mathbf{z}_{-n}\,;\,\nu^*)}\left[\log p(z_n \mid \theta) + \log p(x_n \mid z_n, \theta)\right]\right]\right\} \\ &\propto \exp\left\{\mathbb{E}_{q(\theta\,;\,\nu_0^*)}\left[\log p(z_n \mid \theta) + \log p(x_n \mid z_n, \theta\right]\right\}. \end{aligned}$$

Then

$$q(z_n\,;\,\nu^*) = k_{\mathbf{x}}(x_n) \int_\Theta q(\theta\,;\,\nu_0^*(\mathbf{x})) \log p(z_n \mid \theta) + \log p(x_n \mid z_n, \theta) \mathrm{d}\theta, \tag{2}$$

where $k_{\mathbf{x}}(x_n) = \left[\int_{\mathcal{Z}} \int_\Theta q(\theta\,;\,\nu_0^*(\mathbf{x})) \log p(z_n \mid \theta) + \log p(x_n \mid z_n, \theta) \mathrm{d}\theta \mathrm{d}z_n\right]^{-1}$ is a normalizing constant. The R.H.S of Eq. 2 defines an ideal inference function $f_{\mathbf{x}}(x_n)$, in the sense that, given $\mathbf{x}$, we have $x_n = x_m \implies f_{\mathbf{x}}(x_n) = f_{\mathbf{x}}(x_m)$.

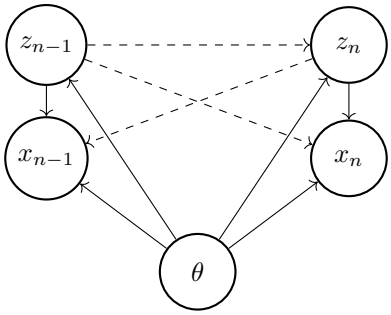

Figure 1: *Graphical representation of a standard latent variable model. If present, the dotted edges preclude the existence of an ideal inference function $f_{\mathbf{x}}(x_n) = \nu_n^*$ and the amortization gap cannot be closed.*

Next we prove the converse, which is item (2) of Theorem 3.3. Applying the CAVI rule to a standard latent variable model,

$$q(z_n\,;\,\nu^*) \propto \exp\left\{\mathbb{E}_{q(\theta,\mathbf{z}_{-n}\,;\,\boldsymbol{\nu}^*_{-n})}\log p(\theta,\mathbf{z},\mathbf{x})\right\}$$
$$\propto \exp\left\{\mathbb{E}_{q(\theta,\mathbf{z}_{-n}\,;\,\boldsymbol{\nu}^*_{-n})}\log p(z_n \mid \mathbf{z}_{-n},\theta) + \log p(x_n \mid z_n,\mathbf{z}_{-n},\theta) + \sum_{i\neq n}\log p(x_i \mid z_n,\mathbf{z}_{-n},\theta)\right\}. \quad (3)$$

The last equation highlights all the terms in which z_n appears. Furthermore, we used the property of *conditional independence* (Definition 3.2 (ii)) to break up the log likelihood $\log p(\mathbf{x} \mid \mathbf{z},\theta)$ into a sum.

Suppose now that there exists a graph $\mathcal{G}$, such that for *any* standard latent variable model supported by this graph, there exists an ideal inference function, that is $\nu_n^* = f_{\mathbf{x}}(x_n)$. Because the q is parametric, we have that the R.H.S of Eq. 3 is also a (dataset dependent) function of x_n. For this assumption to hold *for any choice of distribution*, any contribution of $x_{i\neq n}$ that is not common to all the variational factors of $q(\mathbf{z})$ must be absorbed into the normalizing constant and effectively vanish. We will complete the proof by removing unique contributions of x_i and severing offending edges in $\mathcal{G}$ (Figure 1).

The most obvious contribution of x_i appears in the likelihood terms and is removed if and only if we exclude non-local dependence, that is for $i \neq n$, $p(x_i \mid z_n,\mathbf{z}_{-n},\theta) = p(x_i \mid \mathbf{z}_{-n},\theta)$. Doing so for every n, we have

$$p(x_i \mid z_n,\mathbf{z}_{-n},\theta) = p(x_i \mid z_i,\theta). \quad (4)$$

Remark A.1. Here the assumption of *local dependence* (Definition 3.2 (i)) is critical. Without it, we cannot exclude the possibility that x_i does not depend on z_i, or any z_j's other than z_n, and hence that $p(x_i \mid z_n,\mathbf{z}_{-n},\theta) = p(x_i \mid z_n,\theta)$, $i \neq n$. Then an edge between z_n and x_i would not contradict the existence of an ideal inference function.

Next, we have by assumption that $\nu_i^* = f_{\mathbf{x}}(x_i)$. Then

$$q(z_n\,;\,\nu^*) \propto \exp\left\{\int_{\Theta,\mathcal{Z}_{-n}} q(\mathrm{d}\theta\,;\,\nu_0(\mathbf{x}))\prod_{i\neq n} q(\mathrm{d}z_i\,;\,f_{\mathbf{x}}(x_i))\log p(z_n \mid \mathbf{z}_{-n},\theta) + \log p(x_n \mid z_n,\theta)\right\}. \quad (5)$$

The offending terms are now the variational factors $q(\mathrm{d}z_i\,;\,f_{\mathbf{x}}(x_i))$ in the integral. To remove them, we must get rid of any term that couples z_n and z_i, and so z_n must be a priori independent of z_i, that is

$$p(z_n \mid \mathbf{z}_{-n},\theta) = p(z_n \mid \theta). \quad (6)$$

A standard latent variable model that verifies Eq. 4 and Eq. 6 must also verify Eq. 1 and is therefore a simple hierarchical model. □

A.3 EXAMPLE OF A LATENT VARIABLE MODEL, WHICH IS NOT A SIMPLE HIERARCHICAL MODEL AND ADMITS AN IDEAL INFERENCE FUNCTION

The statement of Theorem 3.4, item (ii) is carefully written for all distributions supported on a graph. To see why a simple "if and only if" version of item (i) is not true, consider a dense hierarchical model, with edges between all elements of $\mathbf{x}$ and

$\mathbf{z}$. If we a choose a likelihood which is symmetric in $\mathbf{z}$, e.g. $p(x_n \mid \mathbf{z}, \theta) = p(x_n \mid \sum_n z_n, \theta)$, then there exists a (constant) ideal inference function and moreover, all factors $q(z_n \,;\, \nu_n^*)$ are identical.

This case is of course trivial: with such a symmetry, the notion of a local latent variable is unjustified. To our knowledge, all examples of models, which are not simple hierarchical models and still admit an ideal inference function, rely on a similar trivialities. These however constitute edge cases we must be mindful of when writing formal statements.

A.4 ANALYTICAL RESULTS FOR THE LINEAR PROBABILISTIC MODEL

We prove Proposition 3.6, which provides an exact expression for the mean and variance of $q(z_n \,;\, \nu^*)$, the optimal solution returned by F-VI when applied to the linear generative model. In the model of interest, θ is a scalar random variable, and we introduce the fixed standard deviations, $\tau \in \mathbb{R}$ and $\sigma \in \mathbb{R}$. Next

$$p(\theta) \propto 1; \;\; p(z_n) = \mathcal{N}(0, 1); \;\; p(x_n) = \mathcal{N}(\theta + \tau z_n, \sigma). \tag{7}$$

Since the posterior distribution $p(\theta, \mathbf{z} \mid \mathbf{x})$ is normal, $q(z_n \,;\, \nu^*)$ can be worked out analytically [e.g Turner and Sahani, 2011, Margossian and Saul, 2023]. Specifically,

$$q(z_n \,;\, \nu_n^*) = \mathcal{N}\left(\mu_n, \frac{1}{[\Sigma^{-1}]_{nn}}\right), \tag{8}$$

where μ_n is the correct posterior mean for z_n and Σ is the correct posterior covariance matrix. Note that F-VI always underestimates the posterior marginal variance unless Σ is diagonal [Margossian and Saul, 2023, Theorem 3.1]. It remains to find an analytical expression for the posterior distribution.

Lemma A.2. *The marginal posterior distribution is given by*

$$p(z_n \mid \mathbf{x}) = \mathcal{N}\left(\frac{\tau}{\sigma^2 + \tau^2}(x_n - \bar{x}), s\right), \tag{9}$$

for some s, constant with respect to $\mathbf{x}$.

Proof. From Bayes' rule

$$\begin{aligned}
\log p(\mathbf{z}, \theta \mid \mathbf{x}) &= k - \frac{1}{2}\sum_{n=1}^N z_n^2 - \frac{1}{2\sigma^2}\sum_{n=1}^N (x_n - \theta - \tau z_n)^2 \\
&= k - \frac{1}{2}\sum_{n=1}^N z_n^2 - \frac{1}{2\sigma^2}\sum_{n=1}^N \theta^2 + (x - \tau z_n)^2 - 2\theta(x_n - \tau z_n) \\
&= k - \frac{1}{2}\sum_{n=1}^N z_n^2 - \frac{1}{2\sigma^2}\left(n\theta^2 + \sum_{n=1}^N (x_n - \tau z_n)^2 - 2\theta\sum_{n=1}^N (x_n - \tau z_n)\right),
\end{aligned} \tag{10}$$

where k is a constant with respect to $\mathbf{z}$ and θ. Moving forward, we overload the notation for k to designate any such constant. As expected, Eq. 10 is quadratic in θ and $\mathbf{z}$.

Remark A.3. At this point, the proof may take two directions: in one, we work out the precision matrix, Φ (i.e. the inverse covariance matrix Σ) for $p(\mathbf{z}, \theta \mid \mathbf{x})$ and invert it to obtain the posterior mean for each z_n. Constructing Φ is straightforward and necessary to show the covariance of $q(z_n \,;\, \nu_n^*)$ is constant with respect to $\mathbf{x}$. However, inverting Φ requires recursively applying the Sherman-Morrison formula three times, which is algebraically tedious. The other direction is to marginalize out θ. We can then construct the precision matrix Ψ for $p(\mathbf{z} \mid \mathbf{x})$, which only requires a single application of the Sherman-Morrison formula to invert. We opt for the second direction, noting both options are rather involved.

To marginalize out θ, we complete the square and perform a Gaussian integral,

$$\begin{aligned}\log p(\mathbf{z}, \theta \mid \mathbf{x}) &= k - \frac{1}{2}\sum_{n=1}^{N} z_n^2 - \frac{n}{2\sigma^2}\left[\theta^2 + \frac{1}{n}\sum_{n=1}^{N}(x_n - \tau z_n)^2 - 2\theta\sum_{n=1}^{N}(x_n - \tau z_n)\right. \\ &\quad \left. + \left(\frac{1}{n}\sum_{n=1}^{N}(x_n - \tau z_n)\right)^2 - \left(\frac{1}{n}\sum_{n=1}^{N}(x_n - \tau z_n)\right)^2\right] \\ &= k - \frac{1}{2}\sum_{n=1}^{N} z_n^2 - \frac{n}{2\sigma^2}\left[\left(\theta - \frac{1}{n}\sum_{n=1}^{N}(x_n - \tau z_n)\right)^2 + \frac{1}{n}\sum_{n=1}^{N}(x_n - \tau z_n)^2\right. \\ &\quad \left. - \left(\frac{1}{n}\sum_{n=1}^{N}(x_n - \tau z_n)\right)^2\right] \end{aligned} \tag{11}$$

Then

$$\log p(\mathbf{z} \mid \mathbf{x}) = k - \frac{1}{2}\sum_{n=1}^{N} z_n^2 - \frac{1}{2\sigma^2}\left[\sum_{n=1}^{N}(x_n - \tau z_n)^2 - \frac{1}{n}\left(\sum_{n=1}^{N}(x_n - \tau z_n)\right)^2\right]. \tag{12}$$

Expanding the square,

$$\left(\sum_{n=1}^{N}(x_n - \tau z_n)\right)^2 = \sum_{n=1}^{N}(x_n - \tau z_n)^2 + 2\sum_{j<n}(x_n - \tau z_n)(x_j - \tau z_j). \tag{13}$$

Plugging this in and factoring out τ, we get

$$\log p(\mathbf{z} \mid \mathbf{x}) = k - \frac{1}{2}\sum_{n=1}^{N} z_n^2 - \frac{\tau^2}{2\sigma^2}\left[\sum_{n=1}^{N}\left(1 - \frac{1}{n}\right)\left(\frac{x_n}{\tau} - z_n\right)^2 - \frac{2}{n}\sum_{j<n}\left(\frac{x_n}{\tau} - z_n\right)\left(\frac{x_j}{\tau} - z_j\right)\right]. \tag{14}$$

Now the standard expression for a multivariate Gaussian is

$$\log p(\mathbf{z} \mid \mathbf{x}) = k - \frac{1}{2}(\mathbf{z} - \boldsymbol{\mu})^T\Psi(\mathbf{z} - \boldsymbol{\mu}) = k - \frac{1}{2}\left(\sum_{n=1}^{N}\Psi_{nn}(z_n - \mu_n)^2 + 2\sum_{j<n}\Psi_{jn}(z_n - \mu_n)(z_j - \mu_j)\right), \tag{15}$$

where $\boldsymbol{\mu}$ is the mean and Ψ the precision matrix. We solve for the mean and precision matrix by matching the coefficients in the above two expressions for z_n, $z_n z_j$, and z_n^2, which respectively produce the following equations:

$$\sum_{j=1}^{N}\Psi_{nj}\mu_j = \frac{\tau}{\sigma^2}(x_n - \bar{x}) \tag{16}$$

$$\Psi_{nj} = -\frac{\tau^2}{n\sigma^2}, \quad \forall n \neq j \tag{17}$$

$$\Psi_{nn} = 1 + \frac{\tau^2}{\sigma^2}\left(1 - \frac{1}{N}\right). \tag{18}$$

This immediately gives us the precision matrix. Eq. 16 may be rewritten in matrix form as

$$\boldsymbol{\mu} = \frac{\tau}{\sigma^2}\Psi^{-1}[\mathbf{x} - \bar{x}\mathbf{1}], \tag{19}$$

where $\mathbf{1}$ is the N-vector of 1's. Let $\alpha = \Psi_{nj}$, for any $n \neq j$, and $\beta = \Psi_{nn} - \alpha$. Then

$$\Psi = \beta I + \alpha\mathbf{1}\mathbf{1}^T, \tag{20}$$

Applying the Sherman-Morrison formula, we obtain the covariance matrix,

$$\begin{aligned}\Psi^{-1} &= (\beta I + \alpha \mathbf{1}\mathbf{1}^T)^{-1} \\ &= \beta^{-1} I - \frac{\beta^{-1} I \alpha \mathbf{1}\mathbf{1}^T \beta^{-1} I}{1 + \alpha \mathbf{1}^T \beta^{-1} I \mathbf{1}} \\ &= \beta^{-1} I - \frac{\alpha \beta^{-1}}{\beta + N\alpha} \mathbf{1}\mathbf{1}^T. \end{aligned} \tag{21}$$

Notice that Ψ^{-1} does not depend on $\mathbf{x}$ and that it's diagonal elements are all equal. Moreover $(\Psi^{-1})_{nn}$ gives us the constant, s. Next let

$$a = \beta^{-1} \frac{\tau}{\sigma^2}; \quad b = -\frac{\alpha \beta^{-1}}{\beta + N\alpha} \frac{\tau}{\sigma^2}. \tag{22}$$

Then $\boldsymbol{\mu} = (aI + b\mathbf{1}\mathbf{1}^T)[\mathbf{x} - \bar{x}\mathbf{1}\mathbf{1}^T]$ and moreover

$$\begin{aligned}\mu_n &= a(x_n - \bar{x}) + b \sum_{j=1}^{N} x_j - \bar{x} \\ &= a(x_n - \bar{x}) \\ &= \frac{\tau}{\sigma^2} \left(\frac{\tau^2 + \sigma^2}{\sigma^2} \right)^{-1} (x_n - \bar{x}) \\ &= \frac{\tau}{\sigma^2 + \tau^2} (x_n - \bar{x}), \end{aligned}$$

as desired.

□

To complete the proof of Proposition 3.4, we need to show that the variances of $q(z_n\,;\,\nu^*)$ is constant with respect to $\mathbf{x}$; that they are equal for each z_n follows from the symmetry of the problem. We already constructed the precision matrix Ψ for $p(\mathbf{z} \mid \mathbf{x})$, but we actually need to study the full precision matrix Φ of $p(\theta, \mathbf{z} \mid \mathbf{x})$. We use the index 0 to denote the columns (or rows) corresponding to θ.

Lemma A.4. *The posterior precision matrix Φ of $p(\theta, \mathbf{z} \mid \mathbf{x})$ verfies*

$$\Phi_{00} = \frac{N}{\sigma^2}; \;\; \Phi_{0j} = \frac{\tau}{2\sigma^2} \text{ if } j > 0; \;\; \Phi_{nn} = 1 + \frac{\tau^2}{\sigma^2} \text{ if } i > 0; \;\; \Phi_{nj} = 0, \text{ if } n \neq j. \tag{23}$$

Crucially, Φ is constant with respect to $\mathbf{x}$.

Proof. Consider Eq. 10, rewritten here for convenience,

$$\log p(\mathbf{z}, \theta \mid \mathbf{x}) = k - \frac{1}{2} \sum_{n=1}^{N} z_n^2 - \frac{1}{2\sigma^2} \left(N\theta^2 + \sum_{n=1}^{N} (x_n - \tau z_n)^2 - 2\theta \sum_{n=1}^{N} (x_n - \tau z_n) \right).$$

The standard Gaussian form is

$$\begin{aligned}\log p(\mathbf{z}, \theta \mid \mathbf{x}) &= k - \frac{1}{2} \Bigg[\Phi_{00}(\theta - \nu)^2 + \sum_{n=1}^{N} \Phi_{nn}(z_n - \mu_n)^2 \\ &\quad + 2 \left(\sum_{j=1}^{N} \Phi_{0j}(\theta - \nu)(z_j - \mu_j) + \sum_{j<n} \Phi_{nj}(z_n - \mu_n)(z_j - \mu_j) \right) \Bigg]. \end{aligned} \tag{24}$$

Matching coefficients for θ^2, θz_j, $z_n z_j$ and z_n^2, we obtain respectively

$$\Phi_{00} = \frac{N}{\sigma^2}; \;\; \Phi_{0j} = \frac{\tau}{2\sigma^2} \text{ if } j > 0; \;\; \Phi_{nn} = 1 + \frac{\tau^2}{\sigma^2} \text{ if } n > 0; \;\; \Phi_{nj} = 0, \text{ if } n \neq j.$$

□

The variance of $q(z_n\,;\,\nu^*)$ is obtained by inverting the diagonal elements of Φ. By symmetry, $\text{Var}_{q^*}(z_n) = \xi \;\; \forall n$, where ξ is a constant which does not depend on $\mathbf{x}$. This completes the proof of Proposition 3.4. □

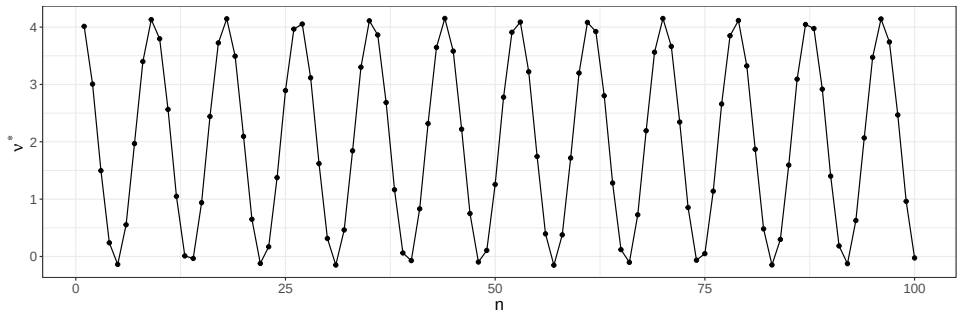

Figure 2: *Optimal variational means when using a Gaussian F-VI on a hidden Markov model (Eq. 25). Even though the elements of* $\mathbf{x}$ *are all equal, the optimal variational means take on different values and so no inference function* $f_\phi : \mathbf{w_n} \to \nu^*_{\mathbf{n}}$ *can be constructed, for any subset* $\mathbf{w_n} \in \mathbf{x}$.

A.5 NON-EXISTENCE OF AN IDEAL INFERENCE FUNCTION FOR HIDDEN MARKOV MODELS

To prove Proposition 3.8, we construct an example for which the optimal F-VI solution, using a factorized Gaussian approximation, can be written in a nearly closed form, and show that the optimal variational factors ν^*_n take different values even when all the values of $\mathbf{x}$ are equal. Then for any strict subset $\mathbf{w}_n \in \mathbf{x}$, we have $\mathbf{w}_n = \mathbf{w}_m$ but $\nu^*_n \neq \nu^*_m$. This provides our counter-example.

Consider the model

$$p(z_0) \propto 1\,;\, p(z_n \mid z_{n-1}) = \mathcal{N}(z_{n-1}, 1)\,;\, p(x_n \mid z_n) = \mathcal{N}(z_n, 1), \tag{25}$$

where θ is held fixed, say to a point estimate $\hat{\theta}$, and ignored for the rest of this analysis. Applying Bayes' rule and expanding

$$\begin{aligned}\log p(\mathbf{z} \mid \mathbf{x}) &= k - \frac{1}{2}\sum_{n=1}^{N}(z_n - z_{n-1})^2 + (x_n - z_n)^2 \\ &= -\frac{1}{2}\sum_{n=1}^{N} 2z_n^2 + z_{n-1}^2 - 2x_n z_n - 2z_n z_{n-1},\end{aligned}$$

which is a quadratic form in $\mathbf{z}$ and hence a Gaussian. Matching the coefficients for z_n, $z_n z_j$ and z_n^2 to the standard expression for a multivariate Gaussian (Eq. 24), we get

$$\sum_{j=1}^{N} \Psi_{nj}\mu_j = -2x_n \tag{26}$$

$$\Psi_{nj} = -2 \quad \text{if } j = n-1 \text{ or } j = n+1 \tag{27}$$

$$\Psi_{nn} = 3 \quad \text{if } n \geq 1 \tag{28}$$

$$\Psi_{00} = 1. \tag{29}$$

All non-specified elements of Ψ go to 0. Moreover the precision matrix Ψ is tri-diagonal. The posterior mean solves the linear problem,

$$\boldsymbol{\mu} = -2\Psi^{-1}\mathbf{x}. \tag{30}$$

Since the variational family and the target are both Gaussian, the optimal variational mean is simply the posterior mean and $\nu^* = \boldsymbol{\mu}$. Even though the elements of $\mathbf{x}$ are all equal, it is in general not the case that the elements of ν^* are constant. To see this explicitly, we take $N = 100$ and $x_1 = x_2 = \cdots = x_N = 1$, and find that the elements of ν^* are indeed distinct (Figure 2). This shows that there exists a hidden Markov model and a realization of the data $\mathbf{x}$ such that no learnable inference function exists. □

B ADDITIONAL EXPERIMENTAL RESULTS

Hardware. All experiments are conducted in `Python` 3.9.15 with `PyTorch` 1.13.1 and `CUDA` 12.0 using an NVIDIA RTX A6000 GPU.

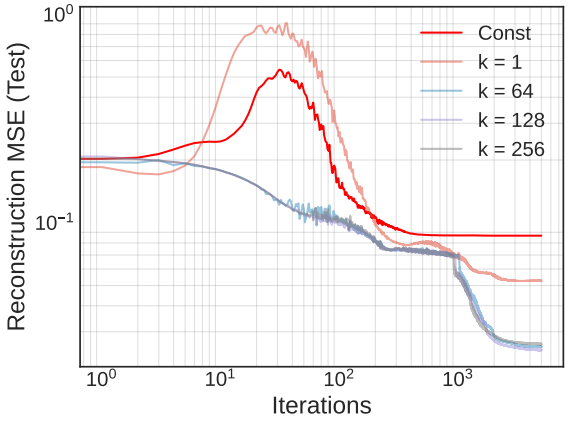

Figure 3: *Reconstruction MSE on a test set.*

Reconstruction error on test set for Bayesian neural network. We consider the reconstruction error on a test set of 10,000 images (Figure 3). The reconstructed image is obtained by (i) computing $q(z' \mid x')$ using the inference function f_ϕ and (ii) feeding $\mathbb{E}_q(z' \mid x')$ into the likelihood neural network Ω (in the VAE context, the "decoder") to obtain $\hat{x}'$. Ω is evaluated at the Bayes estimator $\hat{\theta} = \mathbb{E}_q(\theta \mid \mathbf{x})$. F-VI provides no automatic way of doing step (i) (one would need to learn $q(z' ; \nu')$ by running F-VI from scratch), and so we do not evaluate it on the test set. Overall, we find the model generalizes well, and the test error is very close to the training error.

References

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112, 2017.

Charles C Margossian and Laurence K Saul. The shrinkage-delinkage trade-off: An analysis of factorized gaussian approximations for variational inference. *Uncertainty in Artificial Intelligence*, 2023.

Richard E. Turner and Maneesh Sahani. Two problems with variational expectation maximisation for time-series models. In David Barber, A. Taylan Cemgil, and Silvia Chiappa, editors, *Bayesian Time series models*, chapter 5, pages 109–130. Cambridge University Press, 2011.