Appendix for VFG: Variational Flow Graphical Model with Hierarchical Latent Structure

A Derivation of the ELBO for both Tree and DAG structures

A.1 ELBO of Tree Models

We describe the hierarchical generative network in Figure 7.

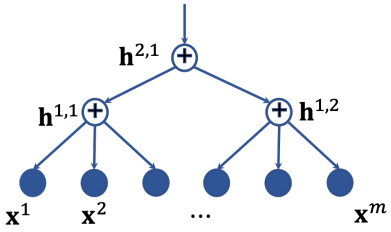


Figure 7: Tree structure.

We assume that for each pair of connected nodes, the edges are invertible mapping functions. The vector of parameters for all the edges is denoted by θ. The forward message passing starts from $\mathbf{x}$ and ends at $\mathbf{h}^L$, and backward message passing in the reverse direction, and Figure 8 illustrates message passing procedure on a tree. Then the likelihood term of the data reads

$$p(\mathbf{x}|\theta) = \sum_{\mathbf{h}^1,\ldots,\mathbf{h}^L} p(\mathbf{h}^L|\theta)p(\mathbf{h}^{L-1}|\mathbf{h}^L,\theta)\cdots p(\mathbf{x}|\mathbf{h}^1,\theta)\,.$$

With the flow-based ensemble model, each edge is invertible. The hierarchical recognition network is the procedure from top to down of the structure as shown in Figure 7. Under independence assumption on the latent nodes, the posterior density of the latent variables is given by

$$q(\mathbf{h}|\mathbf{x},\theta) = q(\mathbf{h}^1|\mathbf{x},\theta)q(\mathbf{h}^2|\mathbf{h}^1,\theta)\cdots q(\mathbf{h}^L|\mathbf{h}^{L-1},\theta)\,,$$

which can be simplified as

$$q(\mathbf{h}|\mathbf{x}) = q(\mathbf{h}^1|\mathbf{x})q(\mathbf{h}^2|\mathbf{h}^1)\cdots q(\mathbf{h}^L|\mathbf{h}^{L-1})\,.$$

Note that we also have

$$q(\mathbf{h}|\mathbf{x}) = q(\mathbf{h}^1|\mathbf{x})q(\mathbf{h}^{2:L}|\mathbf{h}^1)\,. \tag{13}$$

To derive the ELBO of a hierarchical model, we consider all layers of latent variables as the latent vector in conventional VAE, and we have

$$\begin{aligned}
\log p(\mathbf{x}) &= \mathbb{E}_{q(\mathbf{h}|\mathbf{x})}\left[\log \frac{p(\mathbf{x},\mathbf{h})}{p(\mathbf{h}|\mathbf{x})}\right] \\
&= \mathbb{E}_{q(\mathbf{h}|\mathbf{x})}\left[\log \frac{p(\mathbf{x},\mathbf{h})}{q(\mathbf{h}|\mathbf{x})}\frac{q(\mathbf{x},\mathbf{h})}{p(\mathbf{h}|\mathbf{x})}\right] \\
&= \underbrace{\mathbb{E}_{q(\mathbf{h}|\mathbf{x})}\left[\log \frac{p(\mathbf{x},\mathbf{h})}{q(\mathbf{h}|\mathbf{x})}\right]}_{\substack{\mathcal{L}_\theta(x) \\ \text{(ELBO)}}} + \underbrace{\mathbb{E}_{q(\mathbf{h}|\mathbf{x})}\left[\log \frac{q(\mathbf{h}|\mathbf{x})}{p(\mathbf{h}|\mathbf{x})}\right]}_{\mathbf{KL}\left(q(\mathbf{h}|\mathbf{x})|p(\mathbf{h}|\mathbf{x})\right)}\,.
\end{aligned}$$

Since $\mathbf{KL}\big(q(\mathbf{h}|\mathbf{x})|p(\mathbf{h}|\mathbf{x})\big) \geq 0$, as a distance between two distributions, we obtain

$$\begin{aligned}
&\log p(\mathbf{x}) \geq \mathcal{L}_\theta(x) && (14)\\
=&\mathbb{E}_{q(\mathbf{h}|\mathbf{x})}\left[\log \frac{p(\mathbf{x},\mathbf{h})}{q(\mathbf{h}|\mathbf{x})}\right]\\
=&\mathbb{E}_{q(\mathbf{h}^{1:L}|\mathbf{x})}\left[\log \frac{p(\mathbf{x}|\mathbf{h}^{1:L})p(\mathbf{h}^{1:L})}{q(\mathbf{h}^{1:L}|\mathbf{x})}\right]\\
=&\mathbb{E}_{q(\mathbf{h}^{1:L}|\mathbf{x})}\left[\log p(\mathbf{x}|\mathbf{h}^{1:L})\right] + \mathbb{E}_{q(\mathbf{h}^{1:L}|\mathbf{x})}\left[\log \frac{p(\mathbf{h}^{1:L})}{q(\mathbf{h}^{1:L}|\mathbf{x})}\right]\\
=&\mathbb{E}_{q(\mathbf{h}^{1:L}|\mathbf{x})}\left[\log p(\mathbf{x}|\mathbf{h}^{1})\right] + \mathbb{E}_{q(\mathbf{h}^{1:L}|\mathbf{x})}\left[\log \frac{p(\mathbf{h}^{1:L})}{q(\mathbf{h}^{1:L}|\mathbf{x})}\right] && (15)\\
=&\underbrace{\mathbb{E}_{q(\mathbf{h}^{1}|\mathbf{x})}\left[\log p(\mathbf{x}|\mathbf{h}^{1})\right]}_{\substack{\text{Reconstruction of the}\\ \text{data given hidden layer}\\ 1}} + \underbrace{\mathbb{E}_{q(\mathbf{h}^{1:L}|\mathbf{x})}\left[\log \frac{p(\mathbf{h}^{1:L})}{q(\mathbf{h}^{1:L}|\mathbf{x})}\right]}_{-\mathbf{KL}^{1:L}}. && (16)
\end{aligned}$$

The first term in (15) is due to $p(\mathbf{x}|\mathbf{h}^{1:L}) = p(\mathbf{x}|\mathbf{h}^1)$. The first term in (16) is due to the fact that the expectation is regarding $\mathbf{h}^1$. Moreover, the hidden variables $\mathbf{h}^{l+1:L}$ can be taken as the parameters for $\mathbf{h}^l$'s prior distribution . We expand the negated KL term in (16) as follows

$$\begin{aligned}
-\mathbf{KL}^{1:L} =&\mathbb{E}_{q(\mathbf{h}^{1:L}|\mathbf{x})}\left[\log \frac{p(\mathbf{h}^{1:L})}{q(\mathbf{h}^{1:L}|\mathbf{x})}\right] && (17)\\
=&\mathbb{E}_{q(\mathbf{h}^{1:L}|\mathbf{x})}\left[\log \underbrace{\frac{p(\mathbf{h}^{1}|\mathbf{h}^{2:L})p(\mathbf{h}^{2:L})}{q(\mathbf{h}^{1}|\mathbf{x})q(\mathbf{h}^{2:L}|\mathbf{h}^{1})}}_{\text{Due to (13)}}\right]\\
=&\underbrace{\mathbb{E}_{q(\mathbf{h}^{1:L}|\mathbf{x})}\left[\log \frac{p(\mathbf{h}^{1}|\mathbf{h}^{2:L})p(\mathbf{h}^{2:L})}{q(\mathbf{h}^{2:L}|\mathbf{h}^{1})}\right]}_{(a)} + \underbrace{\mathbb{E}_{q(\mathbf{h}^{1:L}|\mathbf{x})}\left[\log \frac{1}{q(\mathbf{h}^{1}|\mathbf{x})}\right]}_{(b)}.
\end{aligned}$$

In forward message passing, the hidden layer $\mathbf{h}^l$ only depends on its previous layer $l-1$. The logarithm term in (a) only relates to hidden states $\mathbf{h}^{1:L}$. With (13), given the hidden states $\mathbf{h}^1$ samples from layer 0, we have

$$(a) = \mathbb{E}_{q(\mathbf{h}^1|\mathbf{x})}\left[\mathbb{E}_{q(\mathbf{h}^{2:L}|\mathbf{h}^1)}\left[\log \frac{p(\mathbf{h}^{1}|\mathbf{h}^{2:L})p(\mathbf{h}^{2:L})}{q(\mathbf{h}^{2:L}|\mathbf{h}^{1})}\right]\right]. \tag{18}$$

The inner expectation is actually the ELBO for hidden variable $\mathbf{h}^1$ of the first layer. Hence

$$\begin{aligned}
&\mathbb{E}_{q(\mathbf{h}^{2:L}|\mathbf{h}^{1})}\left[\log \frac{p(\mathbf{h}^{1}|\mathbf{h}^{2:L})p(\mathbf{h}^{2:L})}{q(\mathbf{h}^{2:L}|\mathbf{h}^{1})}\right]\\
=&\mathbb{E}_{q(\mathbf{h}^{2:L}|\mathbf{h}^{1})}\big[\log p(\mathbf{h}^{1}|\mathbf{h}^{2:L})\big] + \mathbb{E}_{q(\mathbf{h}^{2:L}|\mathbf{h}^{1})}\left[\log \frac{p(\mathbf{h}^{2:L})}{q(\mathbf{h}^{2:L}|\mathbf{h}^{1})}\right]\\
=&\mathbb{E}_{q(\mathbf{h}^{2}|\mathbf{h}^{1})}\big[\log p(\mathbf{h}^{1}|\mathbf{h}^{2})\big] + \mathbb{E}_{q(\mathbf{h}^{2:L}|\mathbf{h}^{1})}\left[\log \frac{p(\mathbf{h}^{2:L})}{q(\mathbf{h}^{2:L}|\mathbf{h}^{1})}\right] && (19)\\
=&\mathbb{E}_{q(\mathbf{h}^{2}|\mathbf{h}^{1})}\big[\log p(\mathbf{h}^{1}|\mathbf{h}^{2})\big] - \mathbf{KL}^{2:L}.
\end{aligned}$$

Term (b) develops as follows:

$$(b) = \mathbb{E}_{q(\mathbf{h}^{1:L}|\mathbf{x})}\left[\log \frac{1}{q(\mathbf{h}^{1}|\mathbf{x})}\right] = \mathbb{E}_{q(\mathbf{h}^{1}|\mathbf{x})}\left[\log \frac{1}{q(\mathbf{h}^{1}|\mathbf{x})}\right] = \mathbf{H}(\mathbf{h}^1|\mathbf{x}). \tag{20}$$

With (17) (18) (19) (20),

$$-\mathbf{KL}^{1:L} = \mathbb{E}_{q(\mathbf{h}^1|\mathbf{x})}\Big[\mathbb{E}_{q(\mathbf{h}^2|\mathbf{h}^1)}\big[\log p(\mathbf{h}^1|\mathbf{h}^2)\big] - \mathbf{KL}^{2:L}\Big] + \mathbf{H}(\mathbf{h}^1|\mathbf{x})\,.$$

Similarly, for layer l, we have

$$\begin{aligned}-\mathbf{KL}^{l:L} =& \mathbb{E}_{q(\mathbf{h}^l|\mathbf{h}^{l-1})}\Big[\mathbb{E}_{q(\mathbf{h}^{l+1}|\mathbf{h}^l)}\big[\log p(\mathbf{h}^l|\mathbf{h}^{l+1})\big] - \mathbf{KL}^{l+1:L}\Big] + \mathbf{H}(\mathbf{h}^l|\mathbf{h}^{l-1}) \\ =& \mathbb{E}_{q(\mathbf{h}^l|\mathbf{h}^{l-1})}\Big[\mathbb{E}_{q(\mathbf{h}^{l+1}|\mathbf{h}^l)}\big[\log p(\mathbf{h}^l|\mathbf{h}^{l+1})\big]\Big] + \mathbf{H}(\mathbf{h}^l|\mathbf{h}^{l-1}) - \mathbf{KL}^{l+1:L}\,.\end{aligned}$$

Given a batch of samples, we compute and store the forward message and the backward message for each node in the forward and backward message passing procedures (Figure 8). The above KL term can be simplified as

$$-\mathbf{KL}^{l:L} = \mathbb{E}_{q(\mathbf{h}^{l+1}|\mathbf{h}^l)}\big[\log p(\mathbf{h}^l|\mathbf{h}^{l+1})\big] + \mathbf{H}(\mathbf{h}^l|\mathbf{h}^{l-1}) - \mathbf{KL}^{l+1:L}\,. \tag{21}$$

For a hierarchical model with L layers, we can recursively expand the KL term, in the ELBO objective, for each layer. Thus

$$\begin{aligned}&\mathbb{E}_{q(\mathbf{h}^{1:L}|\mathbf{x})}\left[\log \frac{p(\mathbf{h}^{1:L})}{q(\mathbf{h}^{1:L}|\mathbf{x})}\right] \\ =&\sum_{l=1}^{L-1}\left\{\mathbb{E}_{q(\mathbf{h}^{l+1}|\mathbf{h}^l)}\Big[\log p(\mathbf{h}^l|\mathbf{h}^{l+1})\Big] + \mathbf{H}(\mathbf{h}^l|\mathbf{h}^{l-1})\right\} \\ &+ \mathbb{E}_{q(\mathbf{h}^L|\mathbf{h}^{L-1})}\Big[\log p(\mathbf{h}^{L-1}|\mathbf{h}^L))\Big] - \mathbf{KL}\big(q(\mathbf{h}^L|\mathbf{h}^{L-1})|p(\mathbf{h}^L)\big)\,.\end{aligned} \tag{22}$$

With $\mathbf{h}^0 = \mathbf{x}$, the ELBO can be expressed as

$$\log p(\mathbf{x}) \geq \sum_{l=0}^{L-1}\mathbb{E}_{q(\mathbf{h}^{l+1}|\mathbf{h}^l)}\Big[\log p(\mathbf{h}^l|\mathbf{h}^{l+1})\Big] + \sum_{l=1}^{L-1}\mathbf{H}(\mathbf{h}^l|\mathbf{h}^{l-1}) - \mathbf{KL}\big(q(\mathbf{h}^L|\mathbf{h}^{L-1})|p(\mathbf{h}^L)\big).$$

The hidden variables are computed with forward message passing with encoders $q(\mathbf{h}^l|\mathbf{h}^{l-1}), l = 1, \ldots, L$. The reconstructed hidden variables are computed with decoders $p(\mathbf{h}^l|\mathbf{h}^{l+1}), l = L-1, \ldots, 0$. We use $\widehat{\mathbf{h}}^l$ to represent the reconstruction of $\mathbf{h}^l$. Only at the root level L, we have $\widehat{\mathbf{h}}^L = \mathbf{h}^L$. Each latent variable is reconstructed with messages from higher layer. Hence the ELBO can be expressed as

$$\log p(\mathbf{x}) \geq \sum_{l=0}^{L-1}\mathbb{E}_{q(\mathbf{h}^{l+1}|\mathbf{h}^l)}\Big[\log p(\mathbf{h}^l|\widehat{\mathbf{h}}^{l+1})\Big] + \sum_{l=1}^{L-1}\mathbf{H}(\mathbf{h}^l|\mathbf{h}^{l-1}) - \mathbf{KL}\big(q(\mathbf{h}^L|\mathbf{h}^{L-1})|p(\mathbf{h}^L)\big)\,.$$

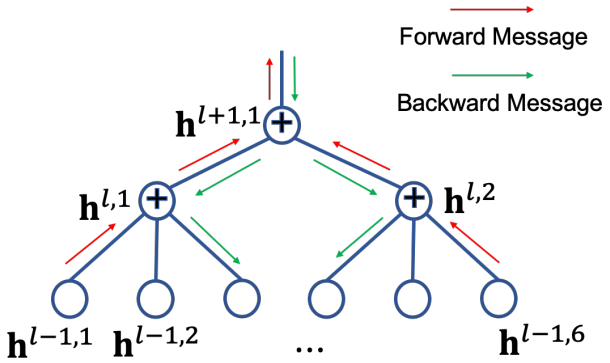


Figure 8: Message passing on a tree.

A.2 ELBO for DAG Models

Note that if we reverse the edge directions in a DAG, the resulting graph is still a DAG graph. The nodes can be listed in a topological order regarding the DAG structure as shown in Figure 9.

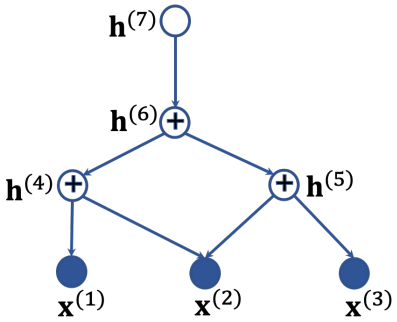


Figure 9: DAG structure. The inverse topology order is { {1,2,3}, {4,5}, {6}, {7} }, and it corresponds to layers 0 to 3.

By taking the topology order as the layers in tree structures, we can derive the ELBO for DAG structures. Assume the DAG structure has L layers, and the root nodes are in layer L. We denote by $\mathbf{h}$ the vector of latent variables, then following (14) we develop the ELBO as

$$\begin{aligned}\log p(\mathbf{x}) \geq \mathcal{L}_\theta(x) =& \mathbb{E}_{q(\mathbf{h}|\mathbf{x})}\left[\log \frac{p(\mathbf{x},\mathbf{h})}{q(\mathbf{h}|\mathbf{x})}\right] \\ =& \underbrace{\mathbb{E}_{q(\mathbf{h}^{pa(\mathbf{x})}|\mathbf{x})}\left[\log p(\mathbf{x}|\mathbf{h}^{pa(\mathbf{x})})\right]}_{\substack{\text{Reconstruction of the data given the parent nodes of the data}}} + \underbrace{\mathbb{E}_{q(\mathbf{h}|\mathbf{x})}\left[\log \frac{p(\mathbf{h})}{q(\mathbf{h}|\mathbf{x})}\right]}_{-\mathbf{KL}} .\end{aligned} \tag{23}$$

Similarly the KL term can be expanded as in the tree structures. For nodes in layer l

$$-\mathbf{KL}^{l:L} = \mathbb{E}_{q(\mathbf{h}^{pa(l)}|\mathbf{h}^l)}\left[\log p(\mathbf{h}^l|\mathbf{h}^{pa(l)})\right] + \mathbf{H}(\mathbf{h}^l|\mathbf{h}^{ch(l)}) - \mathbf{KL}^{l+1:L} . \tag{24}$$

The forward and backward messages or latent state of a node are stored in the message passing procedures. They can be used by the node's parents and children to compute the ELBO. This computation is performed even though the parents or children are not in layer $l+1$ or $l-1$. For the node i in layer l, $pa(i)$ may have children in layers below l. Some nodes in l may not have parent, and combining with the prior, the entropy term will become a KL term in this case. Therefore, we have

$$\begin{aligned}-\mathbf{KL}^{l:L} =& \sum_{i:i\in l, i\notin\mathcal{R}_\mathcal{G}} \left\{\mathbb{E}_{q(\mathbf{h}^{pa(i)}|\mathbf{h}^{ch(pa(i))}}\left[\log p(\mathbf{h}^i|\mathbf{h}^{pa(i)})\right] + \mathbf{H}_q(\mathbf{h}^i|\mathbf{h}^{ch(i)})\right\} \\ &- \sum_{i\in l\bigcap\mathcal{R}_\mathcal{G}} \mathbf{KL}\big(q(\mathbf{h}^{(i)}|\mathbf{h}^{ch(i)})|p(\mathbf{h}^{(i)})\big) - \mathbf{KL}^{l+1:L} .\end{aligned} \tag{25}$$

Recurrently applying (25) yields

$$\begin{aligned}\mathbb{E}_{q(\mathbf{h}|\mathbf{x})}\left[\log \frac{p(\mathbf{h})}{q(\mathbf{h}|\mathbf{x})}\right] =& \sum_{l=1}^{L-1}\sum_{i:i\in l, i\notin\mathcal{R}_\mathcal{G}} \left\{\mathbb{E}_{q(\mathbf{h}^{pa(i)}|\mathbf{h}^{(i)})}\left[\log p(\mathbf{h}^{(i)}|\mathbf{h}^{pa(i)})\right] + \mathbf{H}(\mathbf{h}^i|\mathbf{h}^{ch(i)})\right\} \\ &- \sum_{l=1}^{L-1}\sum_{i\in l\bigcap\mathcal{R}_\mathcal{G}} \mathbf{KL}\big(q(\mathbf{h}^{(i)}|\mathbf{h}^{ch(i)})|p(\mathbf{h}^{(i)})\big) - \mathbf{KL}\big(q(\mathbf{h}^L|\mathbf{h}^{L-1})|p(\mathbf{h}^L)\big) .\end{aligned} \tag{26}$$

Since $L \subseteq \mathcal{R}_{\mathcal{G}}$, with $\mathbf{h}^{(0)} = \mathbf{x}$, (23), and using (26) we have

$$\log p(\mathbf{x}) \geqslant \mathcal{L}(\mathbf{x};\theta) = \sum_{i\in\mathcal{G}\setminus\mathcal{R}_{\mathcal{G}}} \mathbb{E}_{q(\mathbf{h}^{pa(i)}|\mathbf{h}^{ch(pa(i))})}\Big[\log p(\mathbf{h}^{(i)}|\mathbf{h}^{pa(i)})\Big] + \sum_{i\in\mathcal{G}\setminus\mathcal{R}_{\mathcal{G}}} \mathbf{H}(\mathbf{h}^{(i)}|\mathbf{h}^{ch(i)}) - \sum_{i\in\mathcal{R}_{\mathcal{G}}} \mathbf{KL}\big(q(\mathbf{h}^{(i)}|\mathbf{h}^{ch(i)})|p(\mathbf{h}^{(i)})\big)\,.$$

B Theoretical Proofs

We present in this section the proofs for our Lemma 1 and Theorem 1.

B.1 Proof of Lemma 1

Lemma 1. *Let $\mathcal{G}$ be a well trained tree structured variational flow graphical model with L layers, and i and j are two leaf nodes with a as the closest common ancestor. Given observed value at node i, the value of node j can be approximated with $\widehat{\mathbf{x}}^{(j)} \approx \mathbf{f}_{(a,j)}(\mathbf{f}_{(i,a)}(\mathbf{x}^{(i)}))$. Here $\mathbf{f}_{(i,a)}$ is the flow function path from node i to node a. The conditional density of $\mathbf{x}^{(j)}$ given $\mathbf{x}^{(i)}$ can be approximated with*

$$\log p(\mathbf{x}^{(j)}|\mathbf{x}^{(i)}) \approx \log p(\widehat{\mathbf{h}}^{L}) - \frac{1}{2}\log\left(\det\left(\mathbf{J}_{\widehat{\mathbf{x}}^{(j)}}(\widehat{\mathbf{h}}^{L})^{\top}\mathbf{J}_{\widehat{\mathbf{x}}^{(j)}}(\widehat{\mathbf{h}}^{L})\right)\right).$$

Proof. Without loss generality, we assume that there are relationships among different data sections, and the value of one section can be partially or approximately imputed by other sections. According to the aggregation rule (b) discussed in section 3.2, at an aggregation node a, the latent value of a child node j has the same reconstruction value as the parent node. The reconstruction of the child node j can be approximated with the reconstruction of the parent node, i.e., $\widehat{\mathbf{h}}^{(j)} \approx \mathbf{f}_{(a,j)}(\widehat{\mathbf{h}}^{a})$. Recalling the reconstruction term in the ELBO (3), at each node we have $\mathbf{h}^{(a)} \approx \widehat{\mathbf{h}}^{(a)}$. Hence for node a's descendent j, we have $\widehat{\mathbf{h}}^{(j)} \approx \mathbf{f}_{(a,j)}(\mathbf{h}^{(a)})$, and $\mathbf{f}_{(a,j)}$ is the flow function path from a to j. The value of node a can be approximated by the value of its descendent i that has observation, i.e., $\mathbf{h}^{(a)} \approx \mathbf{f}_{(i,a)}(\mathbf{h}^{(i)})$. Hence, we have $\widehat{\mathbf{x}}^{(j)} \approx \mathbf{f}_{(a,j)}(\mathbf{f}_{(i,a)}(\mathbf{x}^{(i)}))$.

To compute node j's conditional distribution given the observed node i, we can use the forward passing to compute the root's reconstruction value $\widehat{\mathbf{h}}^{L}$. Node j's reconstruction value $\widehat{\mathbf{x}}^{(j)}$ can be imputed by backward passing the message at the root. The density value of $\widehat{\mathbf{h}}^{L}$ can be computed with the prior distribution of the root. The conditional density of $\widehat{\mathbf{x}}^{(j)}$ can be computed using the change of variable theorem, and it is known in the context of geometric measure theory as the smooth coarea formula [20, 9]. It reads

$$p(\mathbf{x}^{(j)}|\mathbf{x}^{(i)}) \approx p(\widehat{\mathbf{h}}^{L})\det\left(\mathbf{J}_{\widehat{\mathbf{x}}^{(j)}}(\widehat{\mathbf{h}}^{L})^{\top}\mathbf{J}_{\widehat{\mathbf{x}}^{(j)}}(\widehat{\mathbf{h}}^{L})\right)^{-\frac{1}{2}}.$$

Applying the logarithm operator on both sides concludes the proof of our Lemma.

□

B.2 Proof of Theorem 1

Theorem 1. *Assume that the observed data is distributed according to the model given by (11) and (12). Let the following assumptions holds,*

(a) The sufficient statistics $T_{ij}(h)$ are differentiable almost everywhere and their derivatives $\partial T_{i,j}/\partial_h$ are nonzero almost surely for all $h \in \mathcal{H}_i$, $1 \leq i \leq d$ and $1 \leq j \leq m$.

(b) There exist $(dm+1)$ distinct conditions $\mathbf{u}^{(0)}$, ..., $\mathbf{u}^{(dm)}$ such that the matrix

$$\mathbf{L} = [\lambda(\mathbf{u}^{(1)}) - \lambda(\mathbf{u}^{(0)}), ..., \lambda(\mathbf{u}^{(dm)}) - \lambda(\mathbf{u}^{(0)})]$$

of size $dm \times dm$ is invertible.

Then the model parameters $\mathbf{T}(\mathbf{h}^{(t)}) = \mathbf{A}\widehat{\mathbf{T}}(\mathbf{z}^{(t)}) + \mathbf{c}$. Here $\mathbf{A}$ is a $dm \times dm$ invertible matrix and $\mathbf{c}$ is a vector of size dm.

Proof. The conditional probabilities of $p_{\mathbf{T},\lambda,\mathbf{f}_t^{-1}}\big(\mathbf{x}^{(t)}|\mathbf{u}\big)$ and $p_{\widehat{\mathbf{T}},\widehat{\lambda},\mathbf{g}}\big(\mathbf{x}^{(t)}|\mathbf{u}\big)$ are assumed to be the same in the limit of infinite data. By expanding the probability density functions with the correct change of variable, we have

$$\log p_{\mathbf{T},\lambda}(\mathbf{h}^{(t)}|\mathbf{u}) + \log\big|\det \mathbf{J}_{\mathbf{f}_t}(\mathbf{x}^{(t)})\big| = \log p_{\widehat{\mathbf{T}},\widehat{\lambda}}((\mathbf{h}^{(t)})^\top|\mathbf{u}) + \log\big|\det \mathbf{J}_{g^{-1}}(\mathbf{x}^{(t)})\big|.$$

Let $\mathbf{u}^{(0)}, ..., \mathbf{u}^{(dm)}$ be from condition (b). We can subtract this expression of $\mathbf{u}^{(0)}$ from some $\mathbf{u}^{(v)}$. The Jacobian terms will be removed since they do not depend $\mathbf{u}$,

$$\log p_{\mathbf{h}^{(t)}}(\mathbf{h}^{(t)}|\mathbf{u}^{(v)}) - \log p_{\mathbf{h}^{(t)}}(\mathbf{h}^{(t)}|\mathbf{u}^{(0)}) = \log p_{\mathbf{z}^{(t)}}(\mathbf{z}^{(t)}|\mathbf{u}^{(v)}) - \log p_{\mathbf{z}^{(t)}}(\mathbf{z}^{(t)}|\mathbf{u}^{(0)}). \tag{27}$$

Both conditional distributions in equation 27 belong to the exponential family. Eq. (27) thus reads

$$\begin{aligned}
&\sum_{i=1}^{d}\left[\log\frac{Z_i(\mathbf{u}^{(0)})}{Z_i(\mathbf{u}^{(v)})} + \sum_{j=1}^{m} T_{i,j}(\mathbf{h}^{(t)})\big(\lambda_{i,j}(\mathbf{u}^{(v)}) - \lambda_{i,j}(\mathbf{u}^{(0)})\big)\right] \\
=&\sum_{i=1}^{d}\left[\log\frac{\widehat{Z}_i(\mathbf{u}^{(0)})}{\widehat{Z}_i(\mathbf{u}^{(v)})} + \sum_{j=1}^{m} \widehat{T}_{i,j}(\mathbf{z}^{(t)})\big(\widehat{\lambda}_{i,j}(\mathbf{u}^{(v)}) - \widehat{\lambda}_{i,j}(\mathbf{u}^{(0)})\big)\right].
\end{aligned}$$

Here the base measures Q_is are canceled out. Let $\bar{\lambda}(\mathbf{u}) = \lambda(\mathbf{u}) - \lambda(\mathbf{u}^{(0)})$. The above equation can be expressed, with inner products, as follows

$$\langle \mathbf{T}(\mathbf{h}^{(t)}), \bar{\lambda}\rangle + \sum_i \log\frac{Z_i(\mathbf{u}^{(0)})}{Z_i(\mathbf{u}^{(v)})} = \langle \widehat{\mathbf{T}}(\mathbf{z}^{(t)}), \bar{\widehat{\lambda}}\rangle + \sum_i \log\frac{\widehat{Z}_i(\mathbf{u}^{(0)})}{\widehat{Z}_i(\mathbf{u}^{(v)})}, \quad \forall\ v, 1 \le v \le dm.$$

Combine dm equations together and we can rewrite them in matrix equation form as following

$$\mathbf{L}^\top \mathbf{T}(\mathbf{h}^{(t)}) = \widehat{\mathbf{L}}^\top \widehat{\mathbf{T}}(\mathbf{z}^{(t)}) + \mathbf{b}.$$

Here $b_v = \sum_{i=1}^{d} \log\frac{\widehat{Z}_i(\mathbf{u}^{(0)})Z_i(\mathbf{u}^{(v)})}{\widehat{Z}_i(\mathbf{u}^{(v)})Z_i(\mathbf{u}^{(0)})}$. We can multiply $\mathbf{L}^\top$'s inverse with both sized of the equation,

$$\mathbf{T}(\mathbf{h}^{(t)}) = \mathbf{A}\widehat{\mathbf{T}}(\mathbf{z}^{(t)}) + \mathbf{c}. \tag{28}$$

Here $\mathbf{A} = \mathbf{L}^{-1\top}\widehat{\mathbf{L}}^\top$, and $\mathbf{c} = \mathbf{L}^{-1\top}\mathbf{b}$. By Lemma 1 from [16], there exist m distinct values $h_1^{(t),i}$ to $h_m^{(t),i}$ such that $\big[\frac{dT_i}{dh^{(t),i}}(h_1^{(t),i}), ..., \frac{dT_i}{dh^{(t),i}}(h_m^{(t),i})\big]$ are linearly independent in $\mathbb{R}^m$, for all $1 \le i \le d$. Define m vectors $\mathbf{h}_v^{(t)} = [h_v^{(t),1}, ..., h_v^{(t),d}]$ from points given by this lemma. We obtain the following Jacobian matrix

$$\mathbf{Q} = \left[\mathbf{J}_{\mathbf{T}}(\mathbf{h}_1^{(t)}), ..., \mathbf{J}_{\mathbf{T}}(\mathbf{h}_m^{(t)})\right],$$

where each entry is the Jacobian of size $dm \times d$ from the derivative of Eq. (28) regarding the m vectors $\{\mathbf{h}_j^{(t)}\}_{j=1}^m$. Hence $\mathbf{Q}$ is a $dm \times dm$ invertible by the lemma and the fact that each component of $\mathbf{T}$ is univariate. We can construct a corresponding matrix $\widehat{\mathbf{Q}}$ with the Jacobian of $\widehat{\mathbf{T}}(\mathbf{g}^{-1} \circ \mathbf{f}_t^{-1}(\mathbf{h}^{(t)}))$ computed at the same points and get

$$\mathbf{Q} = \mathbf{A}\widehat{\mathbf{Q}}.$$

Here $\widehat{\mathbf{Q}}$ and $\mathbf{A}$ are both full rank as $\mathbf{Q}$ is full rank. □

According to Theorem 1, the proposed model not only can identify global latent factors, but also identify the latent factors for each section with enough auxiliary information. VFG provides a potential approach to learn the latent hierarchical structures from datasets.

C Additional Numerical Experiments

The flow models employed by VFG in the experiments are implemented with coupling layers [7]. Each block of coupling layer consists of three fully connected layers separated by two RELU layers along with the coupling trick. All latent variables, $\mathbf{h}^i, i \in \mathcal{V}$ are forced to be non-negative via Sigmoid or RELU functions. Non-negativeness can help the model to identify sparse structures of the latent space.

C.1 Latent Representation Learning on MNIST

Figure 10 presents the t-SNE plot of the root latent variables from VFG trained without labels. The figure clearly shows that even without label information, different digits' representation are roughly scattered in different areas. Compared to Figure 6 in section 5.2, label information indeed can improve the latent representation learning.

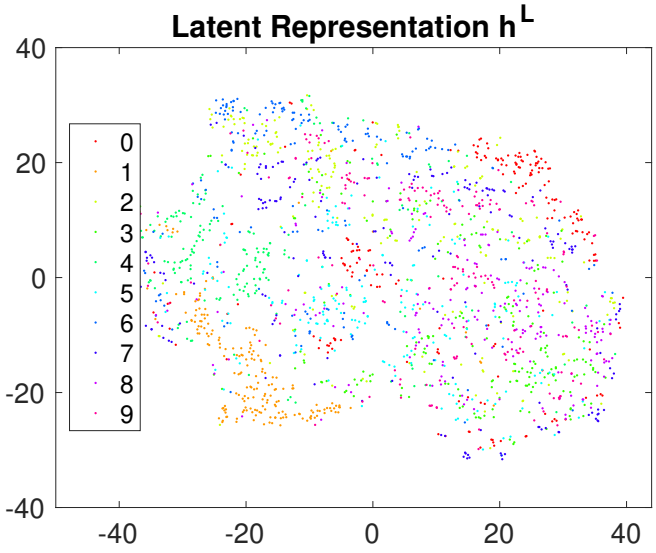
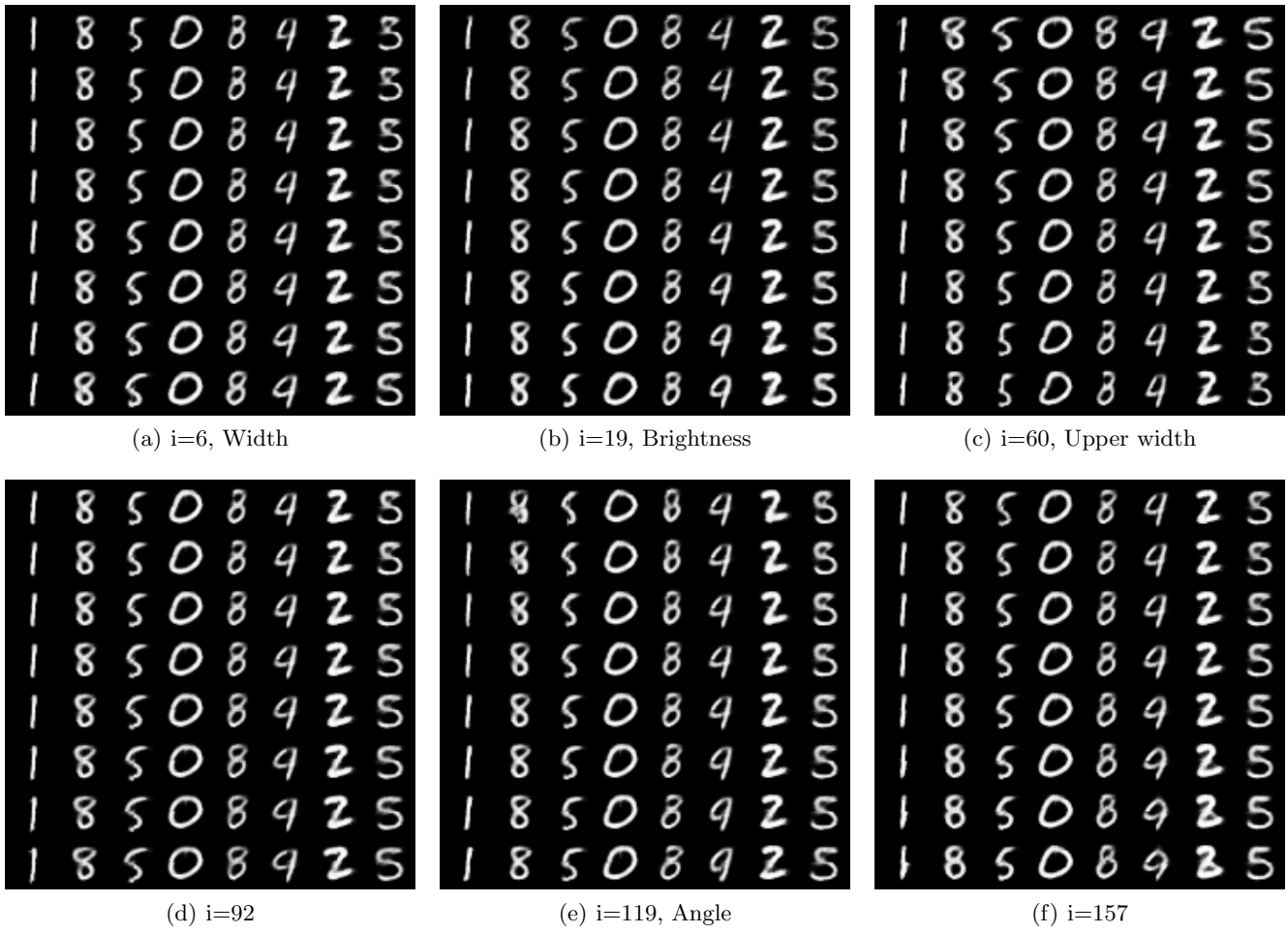


Figure 10: MNIST: t-SNE plot of latent variables from VFG learned without labels.

C.2 Disentanglement on MNIST

We study disentanglement on MNIST with our proposed VFG model introduced in section 5.2. But different from the model in section 5.2, here, the distribution parameter λ for all latent variables are set to be trainable across all layers. Each digit has its trainable vector, $\lambda \in \mathbb{R}^d$ that is used across all layers. To show the disentanglement of learned latent representation, we first obtain the root latent variables of a set of images through forward message passing. Each latent variable's values are changed increasingly within a range centered at the value of the latent variable obtained from the last step. This perturbation is performed for each image in the set. Figure 11 shows the change of images by increasing one latent variable from a small value to a larger one. The figure presents some of the latent variables that have obvious effects on images, and most of the $d = 196$ variables do not impact the generation significantly. Latent variables $i = 6$ and $i = 60$ control the digit width. Variable $i = 19$ affects the brightness. $i = 92, i = 157$ and some of the variables not displayed here control the style of the generated digits.



(a) i=6, Width (b) i=19, Brightness (c) i=60, Upper width

(d) i=92 (e) i=119, Angle (f) i=157

Figure 11: MNIST: Increasing each latent variable from a small value to a larger one.

References

[1] California housing on sklearn. https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_california_housing.html.

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[3] Jeff A Bilmes and Chris Bartels. Graphical model architectures for speech recognition. *IEEE signal processing magazine*, 22(5):89–100, 2005.

[4] Christopher M Bishop, David Spiegelhalter, and John Winn. Vibes: A variational inference engine for bayesian networks. In *Advances in neural information processing systems*, pages 793–800, 2003.

[5] Nicola De Cao, Wilker Aziz, and Ivan Titov. Block neural autoregressive flow. In *Uncertainty in Artificial Intelligence*, pages 1263–1273. PMLR, 2020.

[6] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

[7] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *ArXiv*, abs/1605.08803, 2016.

[8] Bradley Efron et al. Defining the curvature of a statistical problem (with applications to second order efficiency). *The Annals of Statistics*, 3(6):1189–1242, 1975.

[9] Andrew J. Hanson. *Graphics gems iv. chapter Geometry for N-dimensional Graphics*. Academic Press Professional, Inc, 1994.

[10] Jonathan Ho, Xi Chen, Aravind Srinivas, Yan Duan, and Pieter Abbeel. Flow++: Improving flow-based generative models with variational dequantization and architecture design. *arXiv preprint arXiv:1902.00275*, 2019.

[11] 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.

[12] Estevam R Hruschka, Eduardo R Hruschka, and Nelson FF Ebecken. Bayesian networks for imputation in classification problems. *Journal of Intelligent Information Systems*, 29(3):231–252, 2007.

[13] Michael I. Jordan, editor. *Learning in Graphical Models*. MIT Press, Cambridge, MA, USA, 1999.

[14] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[15] David Kahle, Terrance Savitsky, Stephen Schnelle, and Volkan Cevher. Junction tree algorithm. *Stat*, 631, 2008.

[16] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. volume 108 of *Proceedings of Machine Learning Research*, pages 2207–2217, Online, 26–28 Aug 2020. PMLR.

[17] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[18] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10215–10224, 2018.

[19] Daphne Koller, Nir Friedman, Lise Getoor, and Ben Taskar. Graphical models in a nutshell. *Introduction to statistical relational learning*, 43, 2007.

[20] Steven Krantz and Harold Parks. *Analytical Tools: The Area Formula, the Coarea Formula, and Poincaré Inequalities.*, pages 1–33. Birkhäuser Boston, Boston, 2008.

[21] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[22] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances in neural information processing systems*, pages 2378–2386, 2016.

[23] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[24] David Madigan, Jeremy York, and Denis Allard. Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique*, pages 215–232, 1995.

[25] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *arXiv preprint arXiv:1912.02762*, 2019.

[26] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

[27] Oren Rippel and Ryan Prescott Adams. High-dimensional probability estimation with deep density models. *arXiv preprint arXiv:1302.5125*, 2013.

[28] Tim Salimans, Diederik Kingma, and Max Welling. Markov chain monte carlo and variational inference: Bridging the gap. In *International Conference on Machine Learning*, pages 1218–1226, 2015.

[29] Scott Sanner and Ehsan Abbasnejad. Symbolic variable elimination for discrete and continuous graphical models. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[30] Michael Shwe, Blackford Middleton, David Heckerman, Max Henrion, Eric Horvitz, Harold Lehmann, and Gregory Cooper. A probabilistic reformulation of the quick medical reference system. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 790. American Medical Informatics Association, 1990.

[31] Peter Sorrenson, Carsten Rother, and Ullrich Köthe. Disentanglement by nonlinear ica with general incompressible-flow networks (gin). In *Ninth International Conference on Learning Representations*, 2020.

[32] Esteban G Tabak, Eric Vanden-Eijnden, et al. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217–233, 2010.

[33] John Winn and Christopher M Bishop. Variational message passing. *Journal of Machine Learning Research*, 6(Apr):661–694, 2005.

[34] 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.