Variational Inference via a Joint Latent Variable Model with Common Information Extraction

J. Jon Ryu[1], Young-Han Kim[1], Yoojin Choi[2], Mostafa El-Khamy[2], Jungwon Lee[2],
Dept. of ECE, UCSD[1], SoC R&D, Samsung Semiconductor Inc.[2]
{jongharyu,yhk}@ucsd.edu, {yoojin.c,mostafa.e,jungwon2.lee}@samsung.com

UC San Diego JACOBS SCHOOL OF ENGINEERING Electrical and Computer Engineering

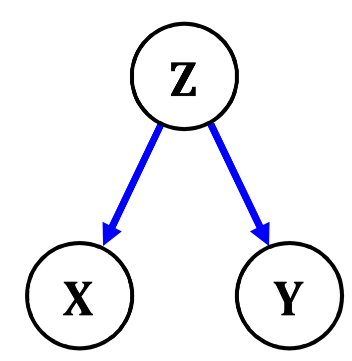
SAMSUNG

Motivation: distributed simulation

- Given $q(\mathbf{x},\mathbf{y})$, two distributed agents wish to generate $\mathbf{X}$ and $\mathbf{Y}$ separately from a shared common randomness and individual local randomnesses
- The least amount of common randomness: Wyner's common information

$$J(\mathbf{X};\mathbf{Y}) = \min_{\mathbf{X}-\mathbf{Z}-\mathbf{Y}} I(\mathbf{X},\mathbf{Y};\mathbf{Z})$$

where the minimum is over all $q(\mathbf{z}|\mathbf{x},\mathbf{y})$ subject to $\mathbf{X}-\mathbf{Z}-\mathbf{Y}$



- We call the minimizer $\mathbf{Z}$ by Wyner's common latent variable
- $I(\mathbf{X},\mathbf{Y};\mathbf{Z})$ naturally quantifies the succinctness of the latent variable $\mathbf{Z}$
- **Our approach**: Use Wyner's common latent variable and the Markov chain $\mathbf{X}-\mathbf{Z}-\mathbf{Y}$ for inference tasks between high-dim. $\mathbf{X}$ and $\mathbf{Y}$

Varitional optimization of Wyner's CI

$$\begin{aligned} &\text{minimize} && I(\mathbf{X}_\theta,\mathbf{Y}_\theta;\mathbf{Z}_\theta)\\ &\text{subject to} && p_\theta(\mathbf{x},\mathbf{y}) = q(\mathbf{x},\mathbf{y})\\ &\text{variables} && p_\theta(\mathbf{z})p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{y}|\mathbf{z})\end{aligned}$$

- With variational bounds and some slackness in the equality constraint:

$$\underset{\theta,\phi}{\text{minimize}}\ D(q(\mathbf{x},\mathbf{y})q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})\|p_\theta(\mathbf{z})p_\theta(\mathbf{x},\mathbf{y}|\mathbf{z})) + \lambda D(q(\mathbf{x},\mathbf{y})q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})\|q(\mathbf{x},\mathbf{y})p_\theta(\mathbf{z}))$$

$$\equiv \underset{\theta,\phi}{\text{minimize}}\ \mathsf{E}_{q(\mathbf{x},\mathbf{y})}\left[(1+\lambda)D(q_\phi(\mathbf{z}|\mathbf{X},\mathbf{Y})\|p_\theta(\mathbf{z})) + \int q_\phi(\mathbf{z}|\mathbf{X},\mathbf{Y})\log\frac{1}{p_\theta(\mathbf{X},\mathbf{Y}|\mathbf{z})}\,d\mathbf{z}\right]$$

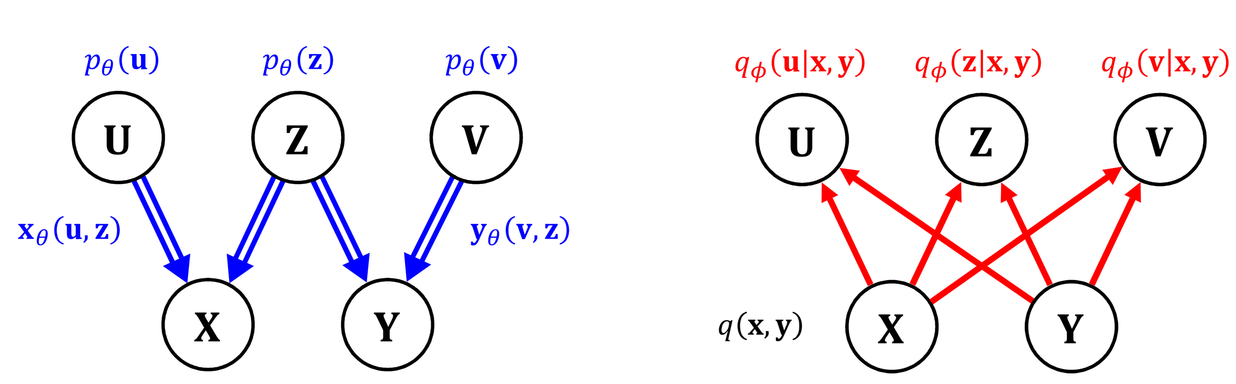
- When $\lambda=0$: boils down to the joint VAE objective in the literature, but still contains $I(\mathbf{X},\mathbf{Y};\mathbf{Z})(=\mathsf{E}_{q(\mathbf{x},\mathbf{y})}[D(q_\phi(\mathbf{z}|\mathbf{X},\mathbf{Y})\|p_\theta(\mathbf{z}))])$

Refined joint latent variable model

- **Reparameterization** of the stochastic decoders

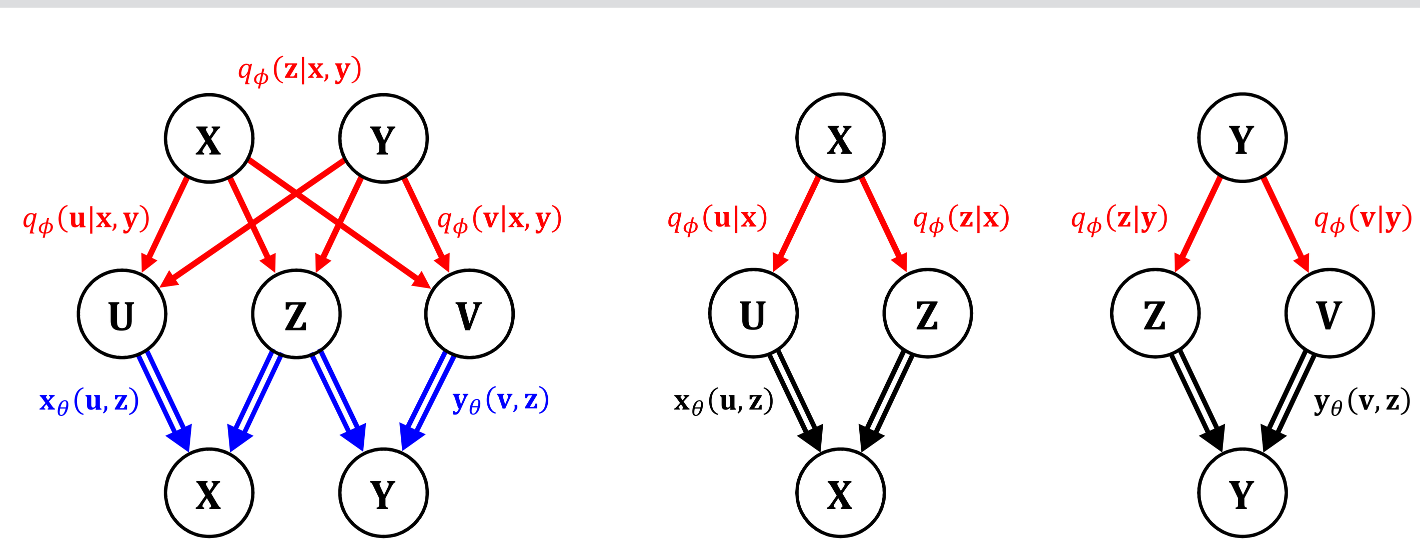
$$p_\theta(\mathbf{x}|\mathbf{z}) = p_\theta(\mathbf{u})\delta(\mathbf{x}-\mathbf{x}_\theta(\mathbf{u},\mathbf{z}))$$

- (+) Increase the expressivity of decoders
- (+) The latent randomness (e.g., $\mathbf{U}$) can be explicitly inferred
- Additionally assume $q_\phi(\mathbf{u},\mathbf{v},\mathbf{z}|\mathbf{x},\mathbf{y}) = q_\phi(\mathbf{u}|\mathbf{x},\mathbf{y})q_\phi(\mathbf{v}|\mathbf{x},\mathbf{y})q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$



(a) $p_\theta(\mathbf{z})p_\theta(\mathbf{u})\delta(\mathbf{x}-\mathbf{x}_\theta(\mathbf{u},\mathbf{z}))p_\theta(\mathbf{v})\delta(\mathbf{y}-\mathbf{y}_\theta(\mathbf{v},\mathbf{z}))$ (b) $q(\mathbf{x},\mathbf{y})q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})q_\phi(\mathbf{u}|\mathbf{x},\mathbf{y})q_\phi(\mathbf{v}|\mathbf{x},\mathbf{y})$

Simple two-step training scheme



(a) Training the joint model (b) Training the marginal models

- Joint model objective

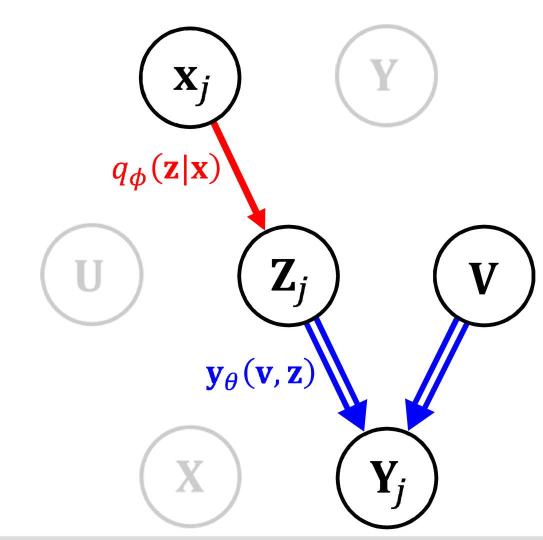
$$\min_\theta \min_\phi D(q_{\text{emp}}(\mathbf{x},\mathbf{y})q_\phi(\mathbf{u},\mathbf{v},\mathbf{z}|\mathbf{x},\mathbf{y})\|p_\theta(\mathbf{u},\mathbf{v},\mathbf{z})\delta(\mathbf{x}-\mathbf{x}_\theta(\mathbf{u},\mathbf{z}))\delta(\mathbf{y}-\mathbf{y}_\theta(\mathbf{v},\mathbf{z})))$$

- For training, delta function $\approx$ Gaussian with small variance

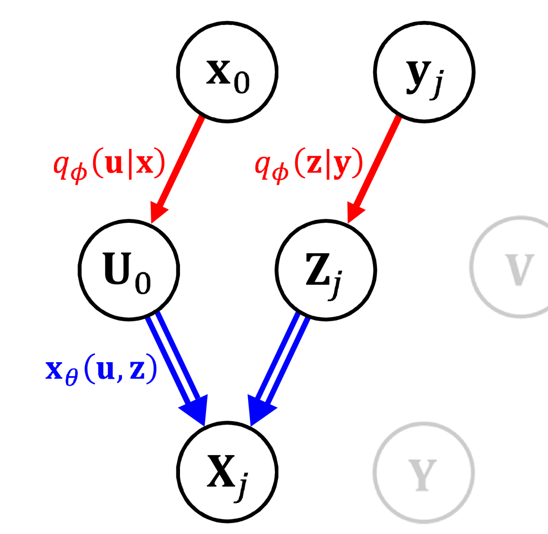
$$\log\frac{1}{\delta(\mathbf{x}-\mathbf{x}_\theta(\mathbf{u},\mathbf{z}))} \approx \frac{1}{2\epsilon^2}\|\mathbf{x}-\mathbf{x}_\theta(\mathbf{u},\mathbf{z})\|^2 + (\text{const.})$$

Applications

(a) Conditional generation.
Genereate $\mathbf{Y}$ given $\mathbf{X}=\mathbf{x}$

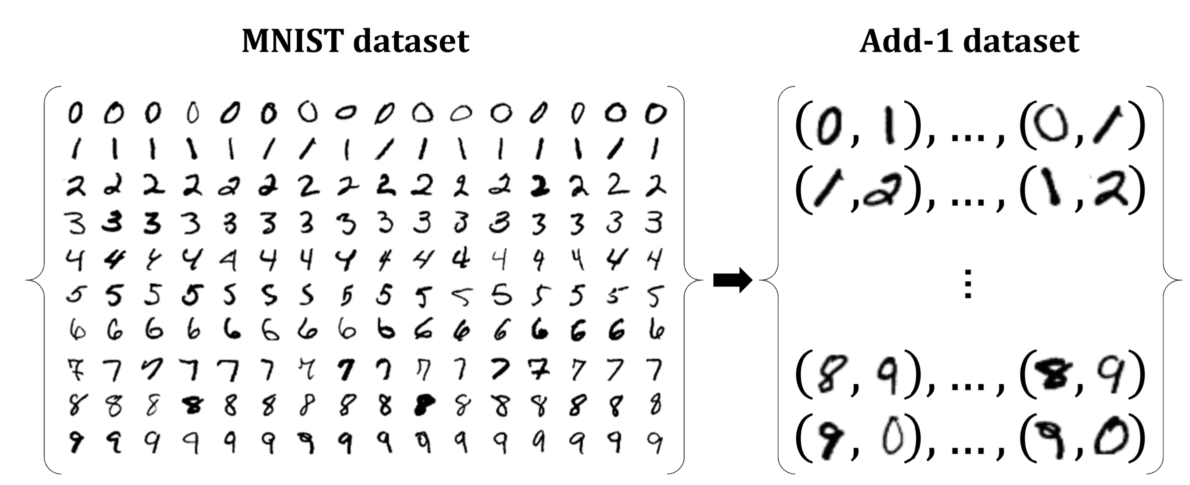


(b) Style transfer.
Generate $\mathbf{X}_j$ conditionally from $\mathbf{y}_j$ in the style of $\mathbf{x}_0$



Experiments

- **Synthetic dataset**



- Take $\lambda=0$ and $\mathbf{Z},\mathbf{U},\mathbf{V}$ as discrete random vectors
- Below: generated multiple $\mathbf{Y}_i$'s given $\mathbf{x}$ for conditional generation

1. Varying $|\mathbf{Z}|$ with fixed $|\mathbf{U}|,|\mathbf{V}|=(N,K)=(2,10)$

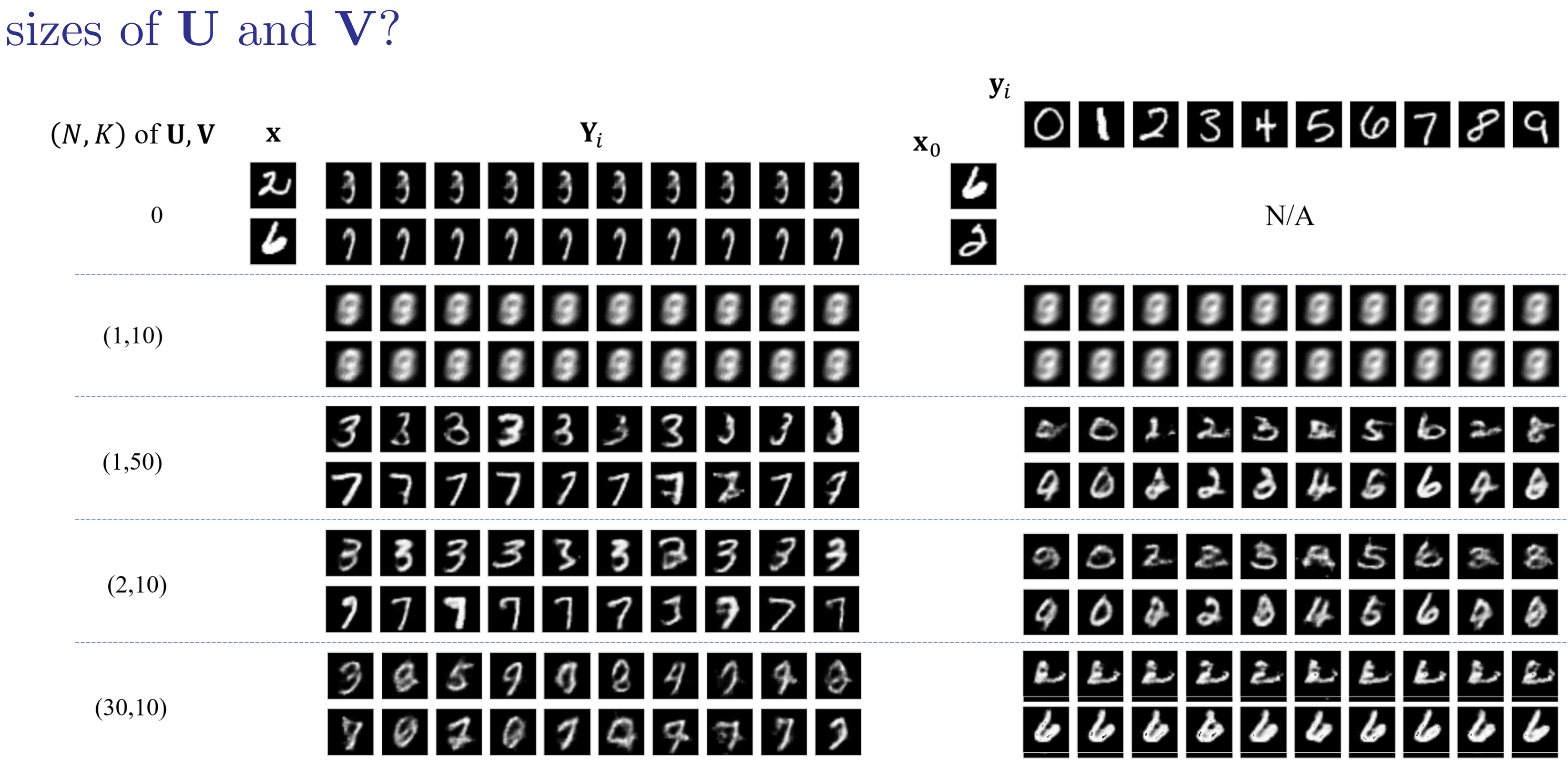
Can we control the amount of common information extraction by manipulating the size of $\mathbf{Z}$?

(a) Conditional generation (b) Style transfer

- Too small $|\mathbf{Z}|$: poor accuracy, i.e., cannot capture label information
- Too large $|\mathbf{Z}|$: poor variability, i.e., results in degenerate style

2. Varying $|\mathbf{U}|,|\mathbf{V}|$ with fixed $|\mathbf{Z}|=(N,K)=(1,100)$

Can we control the amount of expressivity of decoders by manipulating the sizes of $\mathbf{U}$ and $\mathbf{V}$?



(a) Conditional generation (b) Style transfer

- Too small $|\mathbf{U}|,|\mathbf{V}|$: cannot capture style information
- Too large $|\mathbf{U}|,|\mathbf{V}|$: cannot be learned properly