Knowledge-Based Regularization in Generative Modeling

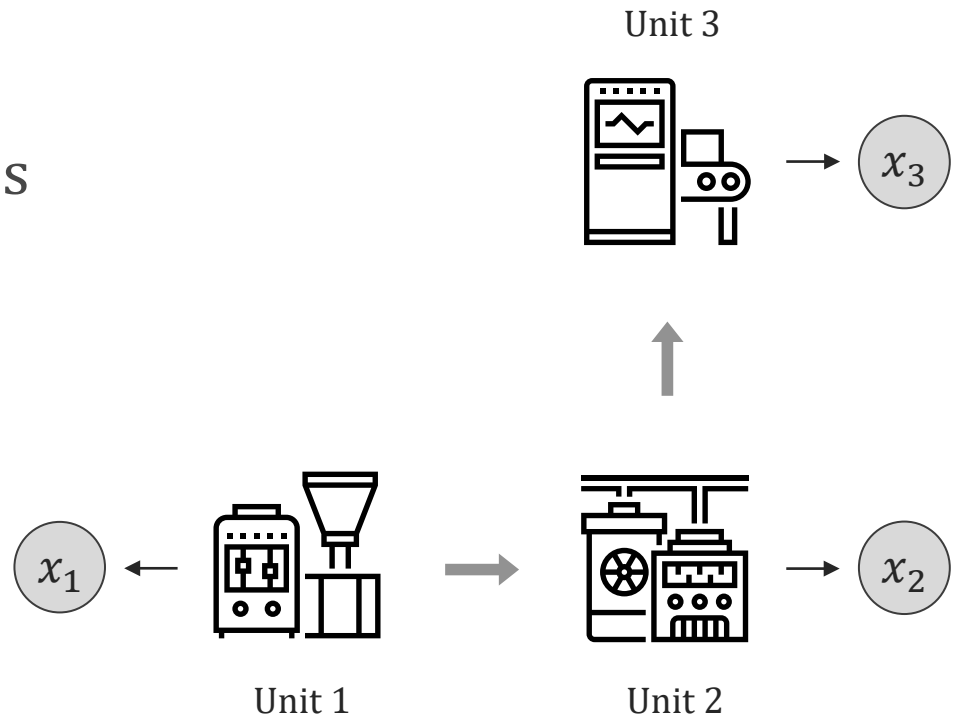
Naoya Takeishi (HES-SO/RIKEN), Yoshinobu Kawahara (KyushuU/RIKEN)

The 29th IJCAI, January 2021

Background

- Generative modeling / density estimation
  - learn $p_\theta(\boldsymbol{x})$ from observations of $\boldsymbol{x}$

- Prior knowledge of data-generating process
  - e.g., system diagram of units and sensors
  - readings of sensors of adjacent units *would be* dependent …
  - … than non-adjacent ones
  - "$\text{dependency}_\theta(x_1, x_2) \geq \text{dependency}_\theta(x_1, x_3)$"
  - Hard-coding such knowledge as model design is costly & bothersome ☹



[drawings created by Eucalyp from the Noun Project]

Proposed method 1/3

- Let $\boldsymbol{x} \in \mathbb{R}^d$ denote a data-point on which we learn generative model $p_\theta(\boldsymbol{x})$
- We express prior knowledge of feature dependence as a *knowledge set* $\mathcal{K}$

Definition (knowledge set).

Let $J \subset \{1, \dots, d\}$ be the index set of d features of $\boldsymbol{x} \in \mathbb{R}^d$. Knowledge of feature dependence is described as a set

$$\mathcal{K} = \left\{ \left(J_s^{\text{ref}},\ J_s^{+},\ J_s^{-} \right) \mid s = 1, \dots, |\mathcal{K}| \right\}.$$

Each triple $\left(J_s^{\text{ref}},\ J_s^{+},\ J_s^{-} \right)$ encodes knowledge that

"$\boldsymbol{x}_{J_s^{\text{ref}}}$ are more dependent on $\boldsymbol{x}_{J_s^{+}}$ than on $\boldsymbol{x}_{J_s^{-}}$."

Proposed method 2/3

$$\underset{\theta}{\text{minimize}} \quad L(\theta) + R_{\mathcal{K}}(\theta)$$

- θ parameters of generative model $p_\theta(x)$
- $L(\theta)$ original loss function for generative modeling (e.g, -ELBO, J-S div.)
- $R_{\mathcal{K}}(\theta)$ proposed regularizer
  - ➢ encourages statistical dependence of marginals of $p_\theta(x)$ to follow prior knowledge

$$R_{\mathcal{K}} = \sum_s \max\left[0,\ \widehat{\text{HSIC}}\left(p_\theta\left(x_{J_s^{\text{ref}}}\right),\ p_\theta\left(x_{J_s^-}\right)\right) - \widehat{\text{HSIC}}\left(p_\theta\left(x_{J_s^{\text{ref}}}\right),\ p_\theta\left(x_{J_s^+}\right)\right)\right]$$

estimated dependency between $p_\theta\left(x_{J_s^{\text{ref}}}\right)$ & $p_\theta\left(x_{J_s^-}\right)$ $\geqslant$ estimated dependency between $p_\theta\left(x_{J_s^{\text{ref}}}\right)$ & $p_\theta\left(x_{J_s^+}\right)$

  - ➢ HSIC can be estimated with samples drawn from $p_\theta(x)$ being learned

Proposed method 3/3

Algorithm (knowledge-regularized gradient method).

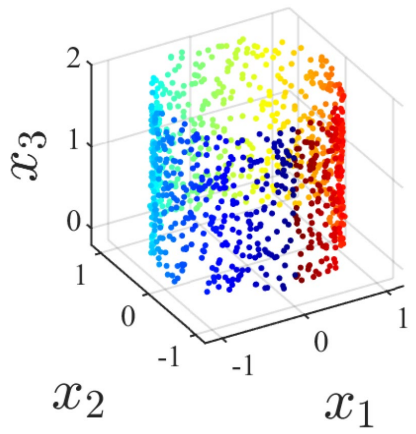
Input: data $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$, knowledge set $\mathcal{K}$, and hyperparameters

Output: parameters θ of a generative model $p_\theta(\boldsymbol{x})$

1. Initialize θ
2. Repeat until convergence:
   a) compute loss L and $\nabla_\theta L$
   b) draw $\hat{\boldsymbol{x}}_1, \dots, \hat{\boldsymbol{x}}_m$ from $p_\theta(\boldsymbol{x})$
   c) for $s = 1, \dots, |\mathcal{K}|$, compute $\widehat{\mathrm{HSIC}}\left(p_\theta\left(x_{J_s^{\mathrm{ref}}}\right), p_\theta\left(x_{J_s^-}\right)\right)$ and $\widehat{\mathrm{HSIC}}\left(p_\theta\left(x_{J_s^{\mathrm{ref}}}\right), p_\theta\left(x_{J_s^+}\right)\right)$
   d) compute regularizer $R_{\mathcal{K}}$ and $\nabla_\theta R_{\mathcal{K}}$
   e) update θ using $\nabla_\theta L$ and $\nabla_\theta R_{\mathcal{K}}$

Experiment: GAN on toy data

Data

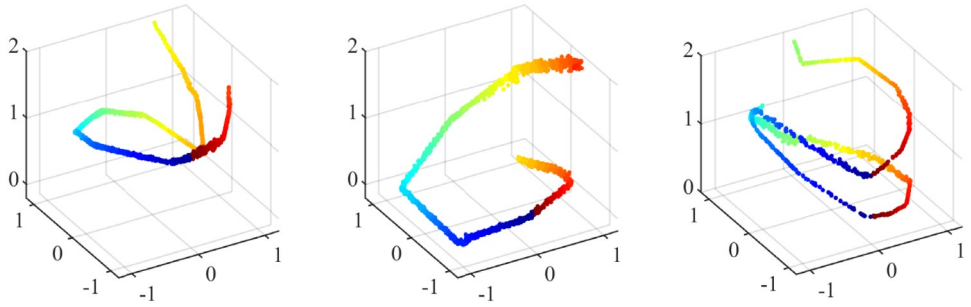


Prior knowledge

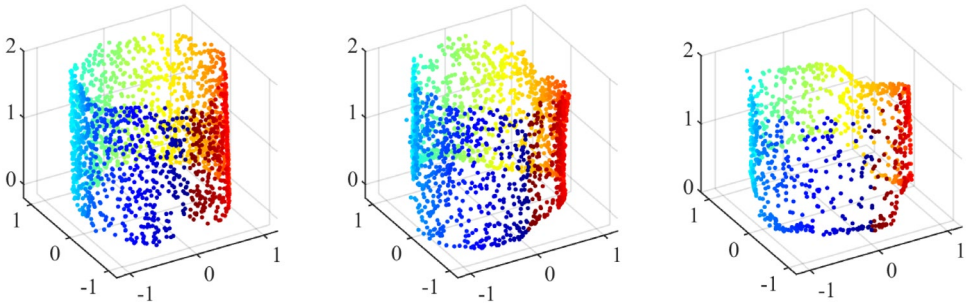
"x_1 & x_2 are more dependent than x_1 & x_3 are"

Result

Samples generated from GAN trained normally (w/o proposed method)



Samples generated from GAN with knowledge (proposed method)



trial #1 trial #2 trial #3

Experiment: VAE on concatenated MNIST images

Data

3 MNIST images concatenated
- top & middle = same type
- bottom = independent type

Prior knowledge

$$\text{dependency}(x_{\text{top}}, x_{\text{middle}}) \geq \text{dependency}(x_{\text{top}}, x_{\text{bottom}})$$

Result

Test avg. cross-entropy of VAEs on ccMNIST dataset (smaller is better)

Setting			Results			
Dataset	dim(z)	dim(MLP)	L2 only	L2 + out-layer	**L2 + proposed**	[dedicated decoder]
ccMNIST	25	512	339.1 (3.6)	339.1 (2.7)	325.1 (2.0) ***	[335.5 (1.7)]
	25	1024	332.0 (3.4)	330.5 (3.6)	320.7 (2.4) ***	[325.2 (1.5)]
	25	2048	328.6 (2.6)	328.5 (2.8)	321.5 (1.5) ***	[324.6 (1.6)]
	50	512	339.0 (4.8)	338.2 (3.9)	321.8 (2.4) ***	[326.1 (3.2)]
	50	1024	325.8 (3.3)	325.3 (3.1)	312.2 (2.6) ***	[300.9 (4.3)]
	50	2048	322.6 (3.5)	321.7 (3.1)	313.8 (2.6) ***	[290.2 (1.7)]
	100	512	339.6 (1.5)	339.6 (1.0)	324.8 (3.4) ***	[326.8 (4.1)]
	100	1024	331.6 (2.9)	329.4 (2.1) *	317.2 (2.0) ***	[304.9 (5.9)]
	100	2048	323.5 (3.6)	323.9 (2.8)	317.4 (2.8) **	[288.9 (2.2)]
	200	512	343.0 (2.6)	340.4 (1.5) **	329.9 (2.2) ***	[328.3 (5.4)]
	200	1024	332.7 (4.4)	332.1 (3.2)	321.5 (2.4) ***	[305.2 (4.2)]
	200	2048	327.2 (1.9)	326.5 (1.8)	324.0 (4.5) *	[290.5 (1.4)]

model configuration

only L2 reg.

regularize similarity of output layer's columns with prior knowledge

proposed method

hard-coding knowledge as model design

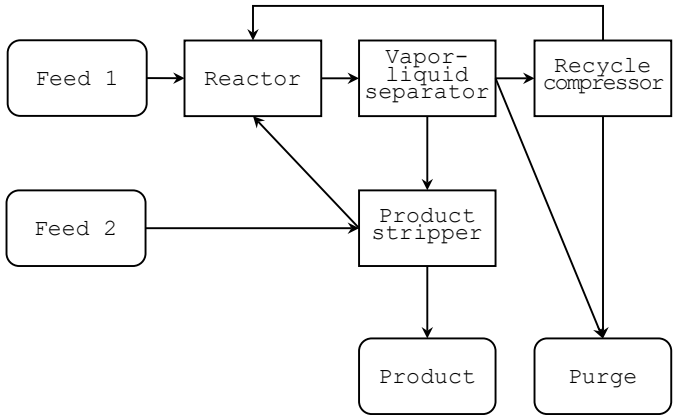
Proposed method achieves intermediate performance between "no knowledge" and "hard-coded knowledge" → trade-off between performance & design cost ☺

Experiment: VAE on plant sensor data

Data

Data comprising readings of 22 sensors of simulated chemical plant (known as Tennessee Eastman process [Downs & Vogel 93])

Prior knowledge



→ adjacent units would be dependent

Result

Test avg. reconstruction error of VAEs on plant sensor dataset (smaller is better)

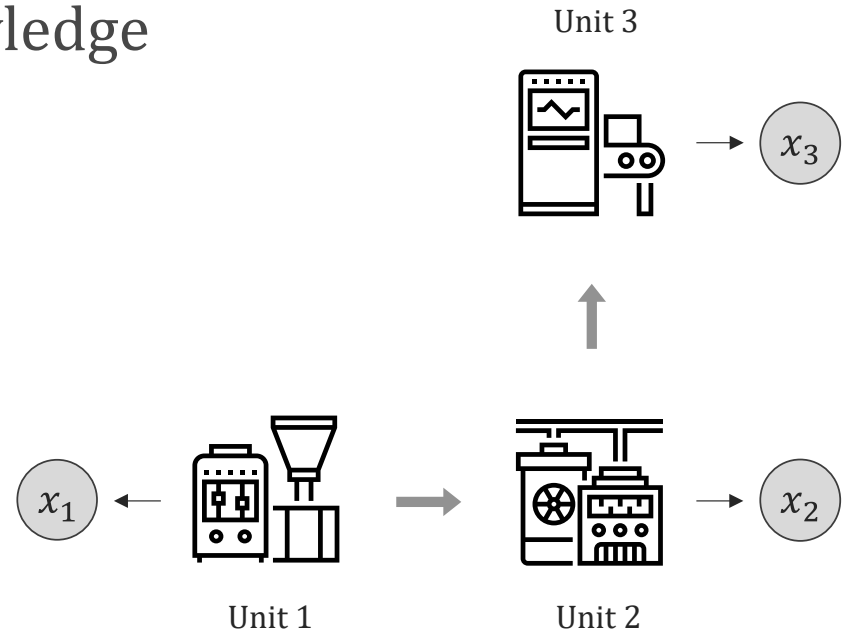
Setting			Results	
Dataset	dim(z)	dim(MLP)	L2 only	**L2 + proposed**
PLANT	4	32	8.776 (0.183)	8.633 (0.112) *
	4	64	8.728 (0.166)	8.672 (0.101)
	4	128	8.778 (0.157)	8.679 (0.104) *
	7	32	8.536 (0.232)	8.322 (0.165) ***
	7	64	8.213 (0.154)	8.061 (0.163) **
	7	128	8.206 (0.073)	8.157 (0.145)
	11	32	8.468 (0.269)	8.374 (0.294)
	11	64	8.106 (0.233)	7.949 (0.166) **
	11	128	7.673 (0.155)	7.567 (0.186) **

model configuration | only L2 reg. | proposed method

Prior knowledge of structure behind data can be incorporated without bothersome model design ☺

Summary

- Learn generative model $p_\theta(x)$ with prior knowledge
  - system diagrams of sensors
  - feature similarities
- Regularize dependency between marginals of $p_\theta(x)$
  - e.g., make $\text{dependency}_\theta(x_1, x_2) \geq \text{dependency}_\theta(x_1, x_3)$
  - using HSIC as dependency measure
- Applicable to VAEs, GANs, etc.
- More results available in arXiv:1902.02068



[drawings created by Eucalyp from the Noun Project]

$$\text{minimize}\ \ R_{\mathcal{K}} = \max\left[0, \widehat{\text{HSIC}}\left(p_\theta(x_1),\ p_\theta(x_3)\right) - \widehat{\text{HSIC}}\left(p_\theta(x_1),\ p_\theta(x_2)\right)\right]$$