PETUUM

SAILING LAB

Laboratory for Statistical Artificial InteLigence & INtegrative Genomics

A Unified View of Deep Generative Models

Zhiting Hu and Eric Xing

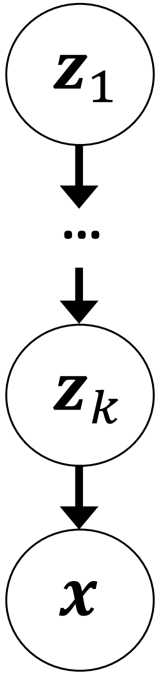
Petuum Inc.

Carnegie Mellon University

Deep generative models

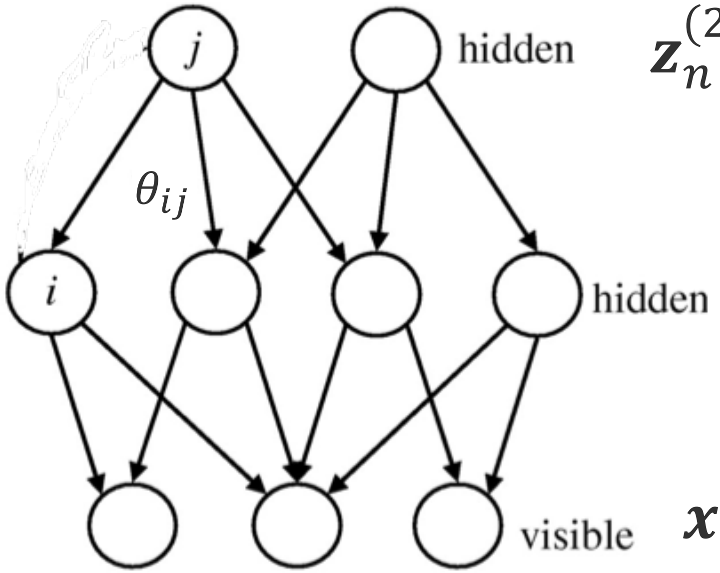
Deep generative models

- Define probabilistic distributions over a set of variables
- "Deep" means multiple layers of hidden variables!



Early forms of deep generative models

- Hierarchical Bayesian models
  - Sigmoid brief nets [Neal 1992



$\boldsymbol{z}_n^{(2)} = \{0,1\}^J$

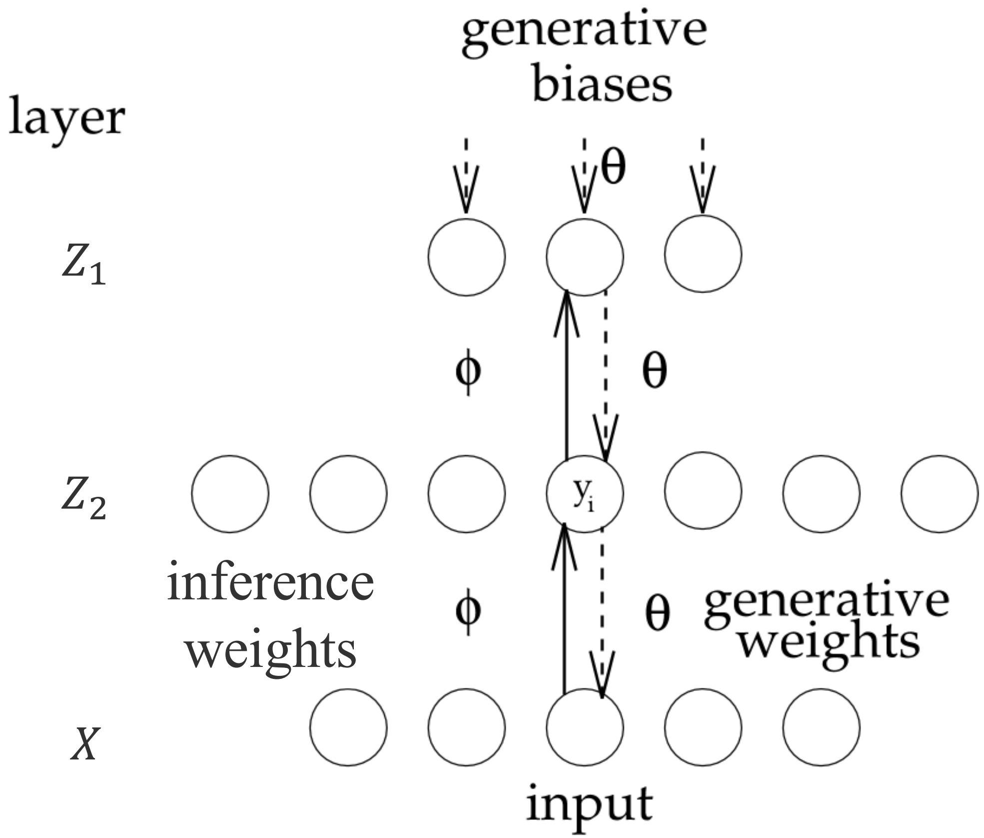
$\boldsymbol{z}_n^{(1)} = \{0,1\}^I$

$\boldsymbol{x}_n = \{0,1\}^K$

$$p\left(x_{kn} = 1 \middle| \boldsymbol{\theta}_k, \boldsymbol{z}_n^{(1)}\right) = \sigma\left(\boldsymbol{\theta}_k^T \boldsymbol{z}_n^{(1)}\right)$$
$$p\left(z_{in}^{(1)} = 1 \middle| \boldsymbol{\theta}_i, \boldsymbol{z}_n^{(2)}\right) = \sigma\left(\boldsymbol{\theta}_i^T \boldsymbol{z}_n^{(2)}\right)$$

Early forms of deep generative models

- Hierarchical Bayesian models
  - Sigmoid brief nets [Neal 1992]
- Neural network models
  - Helmholtz machines [Dayan et al.,1995]



[Dayan et al. 1995]

Early forms of deep generative models

- Hierarchical Bayesian models
  - Sigmoid brief nets [Neal 1992]
- Neural network models
  - Helmholtz machines [Dayan et al.,1995]
  - Predictability minimization [Schmidhuber 1995]

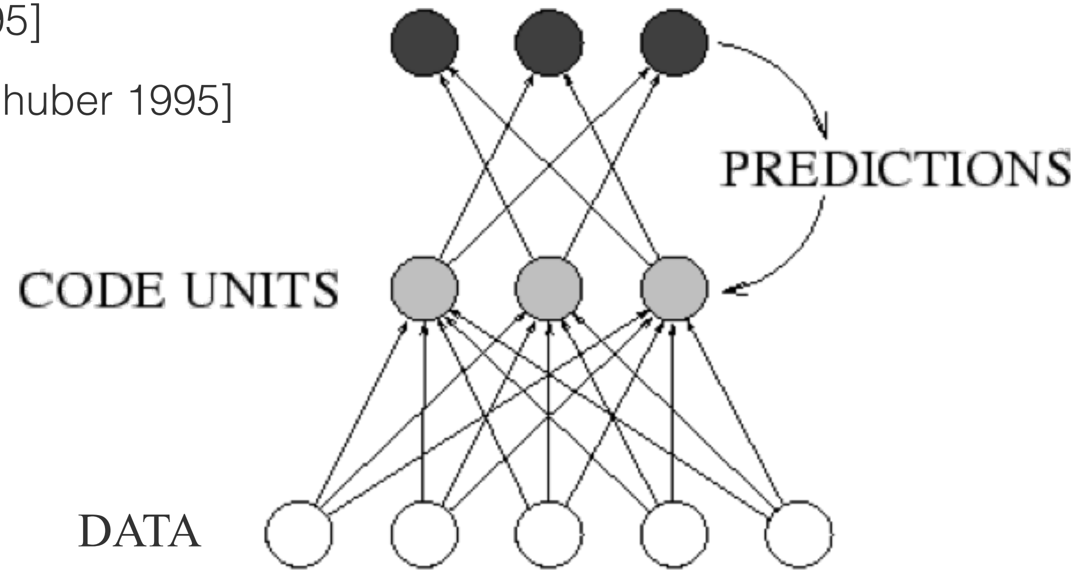


Figure courtesy: Schmidhuber 1996

Early forms of deep generative models

- Training of DGMs via an EM style framework
  - Sampling / data augmentation

$$\boldsymbol{z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2\}$$
$$\boldsymbol{z}_1^{new} \sim p(\boldsymbol{z}_1 | \boldsymbol{z}_2, \boldsymbol{x})$$
$$\boldsymbol{z}_2^{new} \sim p(\boldsymbol{z}_2 | \boldsymbol{z}_1^{new}, \boldsymbol{x})$$

  - Variational inference

$$\log p(\boldsymbol{x}) \geq \mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}, \boldsymbol{z})] - \mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z})) := \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x})$$

$$\max_{\boldsymbol{\theta}, \boldsymbol{\phi}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x})$$

  - Wake sleep

Wake: $\min_\theta \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]$

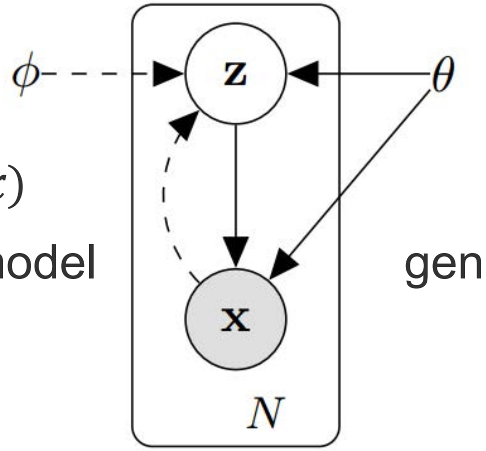
Sleep: $\min_\phi \mathbb{E}_{p_\theta(x|z)}[\log q_\phi(z|x)]$

Resurgence of deep generative models

- Variational autoencoders (VAEs) [Kingma & Welling, 2014]
  / Neural Variational Inference and Learning (NVIL) [Mnih & Gregor, 2014]

$q_\phi(\boldsymbol{z}|\boldsymbol{x})$

inference model



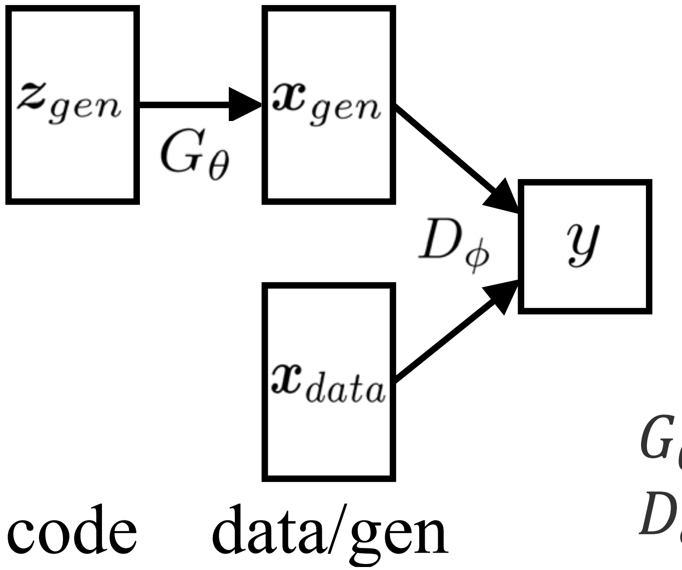
$p_\theta(\boldsymbol{x}|\boldsymbol{z})$

generative model

Figure courtesy: Kingma & Welling, 2014

Resurgence of deep generative models

- Variational autoencoders (VAEs) [Kingma & Welling, 2014]
  / Neural Variational Inference and Learning (NVIL) [Mnih & Gregor, 2014]
- Generative adversarial networks (GANs)



G_θ: generative model
D_ϕ: discriminator

Outline

- Theoretical Basis of deep generative models
  - Wake sleep algorithm
  - Variational autoencoders
  - Generative adversarial networks
- A unified view of deep generative models
  - New formulations of deep generative models
  - Symmetric modeling of latent and visible variables

Synonyms in the literature

- Posterior Distribution -> Inference model
  - Variational approximation
  - Recognition model
  - Inference network (if parameterized as neural networks)
  - Recognition network (if parameterized as neural networks)
  - (Probabilistic) encoder
- "The Model" (prior + conditional, or joint) -> Generative model
  - The (data) likelihood model
  - Generative network (if parameterized as neural networks)
  - Generator
  - (Probabilistic) decoder

Recap: Variational Inference

- Consider a generative model $p_\theta(\boldsymbol{x}|\boldsymbol{z})$, and prior $p(\boldsymbol{z})$
  - Joint distribution: $p_\theta(\boldsymbol{x},\boldsymbol{z}) = p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$
- Assume variational distribution $q_\phi(\boldsymbol{z}|\boldsymbol{x})$
- Objective: Maximize lower bound for log likelihood

$$
\begin{aligned}
&\log p(\boldsymbol{x}) \\
&= KL\left(q_\phi(\boldsymbol{z}|\boldsymbol{x})\,||\,p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right) + \int_{\boldsymbol{z}} q_\phi\,(\boldsymbol{z}|\boldsymbol{x}) \log \frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \\
&\geq \int_{\boldsymbol{z}} q_\phi(\boldsymbol{z}|\boldsymbol{x}) \log \frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q_\phi\,(\boldsymbol{z}|\boldsymbol{x})} \\
&\coloneqq \mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x})
\end{aligned}
$$

  - Equivalently, minimize free energy

$$
F(\boldsymbol{\theta},\phi;\boldsymbol{x}) = -\log p(\boldsymbol{x}) + KL(q_\phi(\boldsymbol{z}|\boldsymbol{x})\,||\,p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}))
$$

Recap: Variational Inference

Maximize the variational lower bound $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x})$

- E-step: maximize $\mathcal{L}$ wrt. ϕ with $\boldsymbol{\theta}$ fixed

$$\max_{\phi} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] + KL(q_\phi\,(z|x)||p(z))$$

  - If with closed form solutions

$$q_\phi^*(z|x) \propto \exp[\log p_\theta(x, z)]$$

- M-step: maximize $\mathcal{L}$ wrt. $\boldsymbol{\theta}$ with ϕ fixed

$$\max_{\theta} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] + KL(q_\phi(z|x)||p(z))$$

Wake Sleep Algorithm

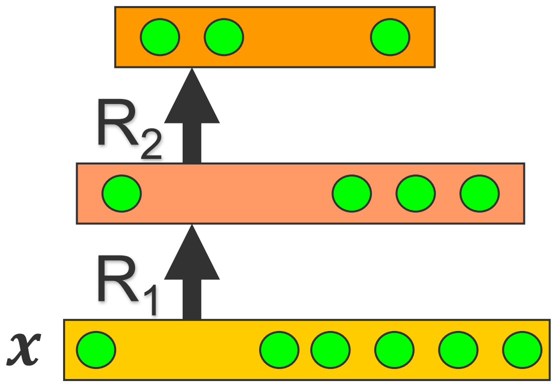
- [Hinton et al., Science 1995]
- Train a separate inference model along with the generative model
  - Generally applicable to a wide range of generative models, e.g., Helmholtz machines
- Consider a generative model $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ and prior $p(\boldsymbol{z})$
  - Joint distribution $p_\theta(\boldsymbol{x},\boldsymbol{z}) = p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$
  - E.g., multi-layer brief nets
- Inference model $q_\phi(\boldsymbol{z}|\boldsymbol{x})$
- Maximize data log-likelihood with two steps of loss relaxation:
  - Maximize the lower bound of log-likelihood, or equivalently, minimize the free energy

$$F(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x}) = -\log p(\boldsymbol{x}) + KL(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}))$$

  - Minimize a different objective (reversed KLD) wrt $\boldsymbol{\phi}$ to ease the optimization
    - Disconnect to the original variational lower bound loss

$$F'(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x}) = -\log p(\boldsymbol{x}) + KL(p_\theta(\boldsymbol{z}|\boldsymbol{x}) \,||\, q_\phi(\boldsymbol{z}|\boldsymbol{x}))$$

Wake Sleep Algorithm



- Free energy:

$$F(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}) = -\log p(\boldsymbol{x}) + KL(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}))$$

- Minimize the free energy wrt. $\boldsymbol{\theta}$ of p_θ → *wake* phase

$$\max_{\boldsymbol{\theta}} \mathrm{E}_{q_{\phi}(\boldsymbol{z}|\boldsymbol{x})} [\log p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})]$$

  - Get samples from $q_\phi(z|x)$ through inference on hidden variables
  - Use the samples as targets for updating the generative model $p_\theta(\boldsymbol{z}|\boldsymbol{x})$
  - Correspond to the variational M step

[Figure courtesy: Maei's slides]

Wake Sleep Algorithm

- Free energy:

$$F(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}) = -\log p(\boldsymbol{x}) + KL(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}))$$

- Minimize the free energy wrt. ϕ of $q_{\phi}(\boldsymbol{z}|\boldsymbol{x})$
  - Correspond to the variational E step
  - Difficulties:
    - Optimal $q_{\boldsymbol{\phi}}^{*}(\boldsymbol{z}|\boldsymbol{x}) = \frac{\boldsymbol{p_\theta(z, x)}}{\int \boldsymbol{p_\theta(z, x)\, dz}}$ intractable
    - High variance of direct gradient estimate $\nabla_{\phi} F(\theta, \phi; x) = \cdots + \nabla_{\phi} \mathbb{E}_{q_{\phi}(z|x)}[\log p_{\theta}(z, x)] + \cdots$
      - Gradient estimate with the log-derivative trick:

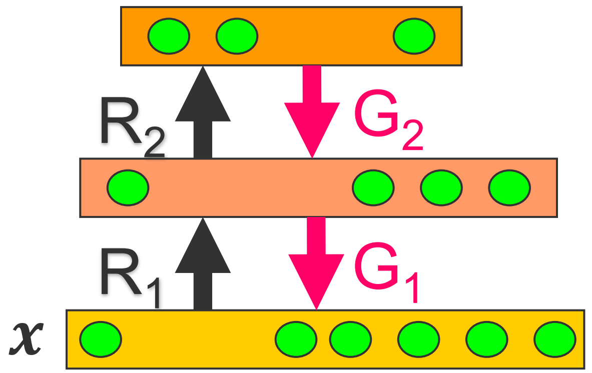
$$\nabla_{\phi} \mathbb{E}_{q_{\phi}}[\log p_{\theta}] = \int \nabla_{\phi} q_{\phi} \log p_{\theta} = \int q_{\phi} \log p_{\theta} \, \nabla_{\phi} \log q_{\phi} = \mathbb{E}_{q_{\phi}}[\log p_{\theta} \, \nabla_{\phi} \log q_{\phi}]$$

        - Monte Carlo estimation:

$$\nabla_{\phi} \mathbb{E}_{q_{\phi}}[\log p_{\theta}] \approx \mathbb{E}_{z_i \sim q_{\phi}}[\log p_{\theta}(x, z_i) \, \nabla_{\phi} q_{\phi}(z_i|x)]$$

        - The scale factor $\log p_{\theta}$ of the derivative $\nabla_{\phi} \log q_{\phi}$ can have arbitrary large magnitude

Wake Sleep Algorithm



- Free energy:

$$F(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}) = -\log p(\boldsymbol{x}) + KL(q_{\boldsymbol{\phi}}(\mathbf{z}|\boldsymbol{x}) \,||\, p_{\boldsymbol{\theta}}(\mathbf{z}|\boldsymbol{x}))$$

- WS works around the difficulties with the sleep phase approximation
- Minimize the following objective → *sleep* phase

$$F'(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}) = -\log p(\boldsymbol{x}) + KL(p_{\theta}(\mathbf{z}|\boldsymbol{x}) \,||\, q_{\phi}(\mathbf{z}|\boldsymbol{x}))$$

$$\max_{\boldsymbol{\phi}} \mathrm{E}_{p_{\theta}(\boldsymbol{z},\boldsymbol{x})}\left[\log q_{\phi}(\mathbf{z}|\boldsymbol{x})\right]$$

  - "Dreaming" up samples from $p_{\theta}(\boldsymbol{x}|\boldsymbol{z})$ through top-down pass
  - Use the samples as targets for updating the inference model
- (Recent approaches other than sleep phase is to reduce the variance of gradient estimate: slides later)

[Figure courtesy: Maei's slides]

Wake Sleep Algorithm

Wake sleep

- Parametrized inference model $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$
- Wake phase:
  - minimize $KL(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}))$ wrt. θ
  - $\mathrm{E}_{q_{\phi}(\boldsymbol{z}|\boldsymbol{x})}\left[\nabla_{\theta} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right]$
- Sleep phase:
  - minimize $KL(p_{\theta}(\boldsymbol{z}|\boldsymbol{x}) \,||\, q_{\phi}(\boldsymbol{z}|\boldsymbol{x}))$ wrt. ϕ
  - $\mathrm{E}_{p_{\theta}(\boldsymbol{z},\boldsymbol{x})}\left[\nabla_{\phi} \log q_{\phi}(\boldsymbol{z},\boldsymbol{x})\right]$
  - low variance
  - Learning with generated samples of $\boldsymbol{x}$
- Two objective, not guaranteed to converge

Variational EM

- Variational distribution $q_{\phi}(\boldsymbol{z}|\boldsymbol{x})$
- Variational M step:
  - minimize $KL(q_{\phi}(\boldsymbol{z}|\boldsymbol{x}) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}))$ wrt. θ
  - $\mathrm{E}_{q_{\phi}(\boldsymbol{z}|\boldsymbol{x})}\left[\nabla_{\theta} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right]$
- Variational E step:
  - minimize $KL(q_{\phi}(\boldsymbol{z}|\boldsymbol{x}) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}))$ wrt. ϕ
  - $q_{\phi}^{*} \propto \exp[\log p_{\theta}]$ if with closed-form
  - $\nabla_{\phi} \mathbb{E}_{q_{\phi}}[\log p_{\theta}(z, x)]$
    - need variance-reduce in practice
  - Learning with real data $\boldsymbol{x}$
- Single objective, guaranteed to converge

Variational Autoencoders (VAEs)

- [Kingma & Welling, 2014]
- Use variational inference with an inference model
  - Enjoy similar applicability with wake-sleep algorithm
- Generative model $p_\theta(\boldsymbol{x}|\boldsymbol{z})$, and prior $p(\boldsymbol{z})$
  - Joint distribution $p_\theta(\boldsymbol{x},\boldsymbol{z}) = p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$
- Inference model $q_\phi(\boldsymbol{z}|\boldsymbol{x})$

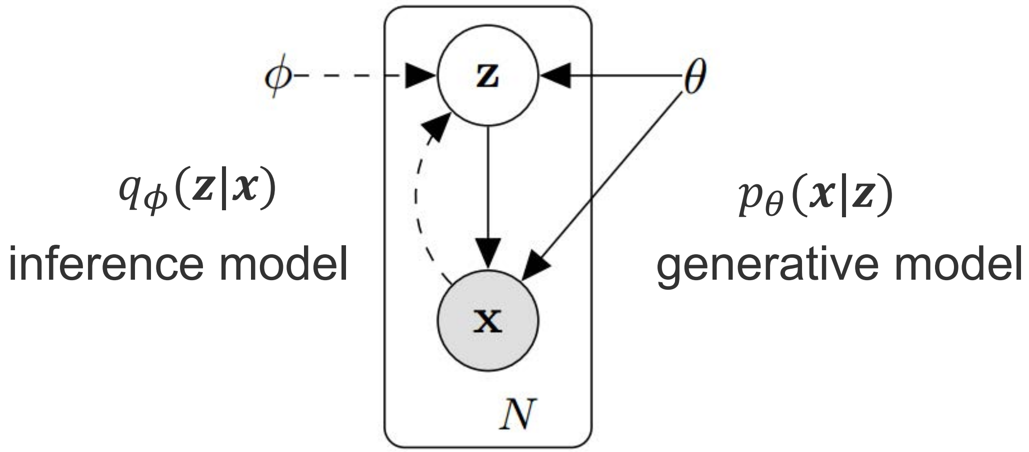


Figure courtesy: Kingma & Welling, 2014

Variational Autoencoders (VAEs)

- Variational lower bound

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}) = \mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}, \boldsymbol{z})] - \mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}))$$

- Optimize $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x})$ wrt. θ of $p_\theta(\boldsymbol{x}|\boldsymbol{z})$
  - The same with the wake phase
- Optimize $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x})$ wrt. ϕ of $q_\phi(\boldsymbol{z}|\boldsymbol{x})$

$$\nabla_\phi \mathcal{L}(\theta, \phi; x) = \cdots + \nabla_\phi \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] + \cdots$$

  - Use *reparameterization trick* to reduce variance
  - Alternatives: use control variates as in reinforcement learning [Mnih & Gregor, 2014; Paisley et al., 2012]

Reparametrized gradient

- Optimize $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x})$ wrt. ϕ of $q_{\phi}(\boldsymbol{z}|\boldsymbol{x})$
  - Recap: gradient estimate with log-derivative trick:

$$\nabla_{\phi}\mathbb{E}_{q_{\phi}}[\log p_{\theta}(\boldsymbol{x}, \boldsymbol{z})] = \mathbb{E}_{q_{\phi}}[\log p_{\theta}(\boldsymbol{x}, \boldsymbol{z})\ \nabla_{\phi}\log q_{\phi}]$$

    - High variance: $\nabla_{\phi}\mathbb{E}_{q_{\phi}}[\log p_{\theta}] \approx \mathbb{E}_{z_{i}\sim q_{\phi}}[\log p_{\theta}(x, z_i)\ \nabla_{\phi} q_{\phi}(z_i|x)]$
      - The scale factor $\log p_{\theta}(x, z_i)$ of the derivative $\nabla_{\phi}\log q_{\phi}$ can have arbitrary large magnitude

  - gradient estimate with *reparameterization trick*

$$\boldsymbol{z} \sim q_{\phi}(\boldsymbol{z}|\boldsymbol{x}) \quad \Leftrightarrow \quad \mathbf{z} = \mathrm{g}_{\phi}(\boldsymbol{\epsilon}, \boldsymbol{x}), \qquad \boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})$$

$$\nabla_{\phi}\mathrm{E}_{q_{\phi}(\boldsymbol{z}|\boldsymbol{x})}[\log p_{\theta}(\boldsymbol{x}, \boldsymbol{z})] = \mathrm{E}_{\epsilon\sim p(\epsilon)}\left[\nabla_{\phi}\log p_{\theta}\left(\boldsymbol{x}, \boldsymbol{z}_{\phi}(\boldsymbol{\epsilon})\right)\right]$$

    - (Empirically) lower variance of the gradient estimate
    - E.g., $\boldsymbol{z} \sim N\left(\boldsymbol{\mu}(\boldsymbol{x}), \boldsymbol{L}(\boldsymbol{x})\boldsymbol{L}(\boldsymbol{x})^{\boldsymbol{T}}\right) \ \Leftrightarrow \ \boldsymbol{\epsilon} \sim N(0,1),\ \boldsymbol{z} = \boldsymbol{\mu}(\boldsymbol{x}) + \boldsymbol{L}(\boldsymbol{x})\boldsymbol{\epsilon}$

VAEs: example results

- VAEs tend to generate blurred images due to the mode covering behavior (more later)

Celebrity faces [Radford 2015]

- Latent code interpolation and sentences generation from VAEs [Bowman et al., 2015].

“ i want to talk to you . ”
“i want to be with you . ”
“i do n’t want to be with you . ”
i do n’t want to be with you .
she did n’t want to be with him .

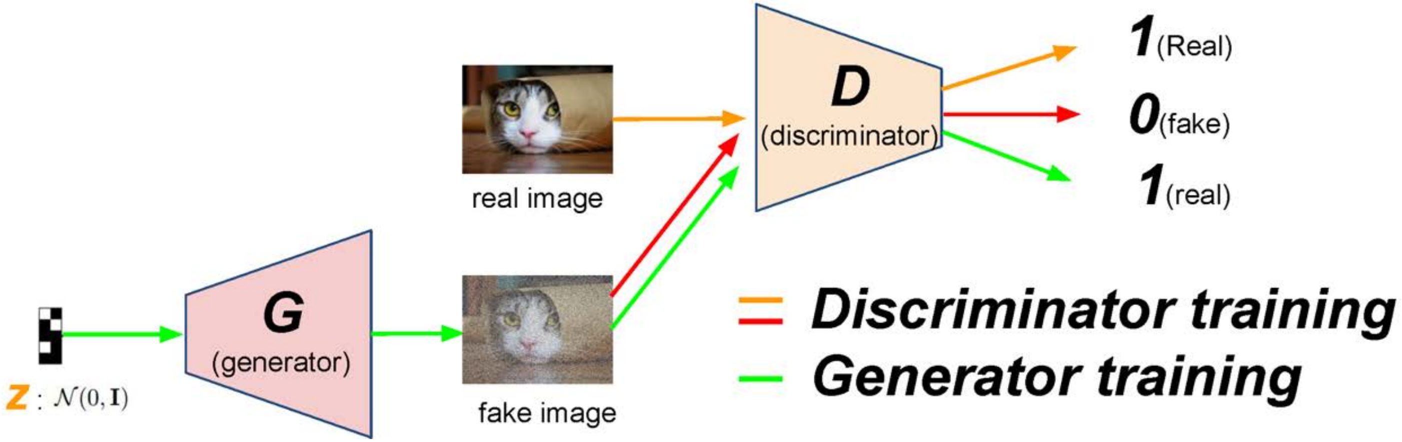
Generative Adversarial Nets (GANs)

- [Goodfellow et al., 2014]
- Generative model $\boldsymbol{x} = G_\theta(\boldsymbol{z}), \quad \boldsymbol{z} \sim p(\boldsymbol{z})$
  - Map noise variable $\boldsymbol{z}$ to data space $\boldsymbol{x}$
  - Define an implicit distribution over $\boldsymbol{x}$: $p_{g_\theta}(\boldsymbol{x})$
    - a stochastic process to simulate data $\boldsymbol{x}$
    - Intractable to evaluate likelihood
- Discriminator $D_\phi(\boldsymbol{x})$
  - Output the probability that $\boldsymbol{x}$ came from the data rather than the generator
- No explicit inference model
- No obvious connection to previous models with inference networks like VAEs
  - We will build formal connections between GANs and VAEs later

Generative Adversarial Nets (GANs)

- Learning
  - A minimax game between the generator and the discriminator
  - Train D to maximize the probability of assigning the correct label to both training examples and generated samples
  - Train G to fool the discriminator

$$\max_D \mathcal{L}_D = \mathbb{E}_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}\left[\log D(\boldsymbol{x})\right] + \mathbb{E}_{\boldsymbol{x}\sim G(\boldsymbol{z}),\boldsymbol{z}\sim p(\boldsymbol{z})}\left[\log(1 - D(\boldsymbol{x}))\right]$$
$$\min_G \mathcal{L}_G = \mathbb{E}_{\boldsymbol{x}\sim G(\boldsymbol{z}),\boldsymbol{z}\sim p(\boldsymbol{z})}\left[\log(1 - D(\boldsymbol{x}))\right].$$



[Figure courtesy: Kim's slides]

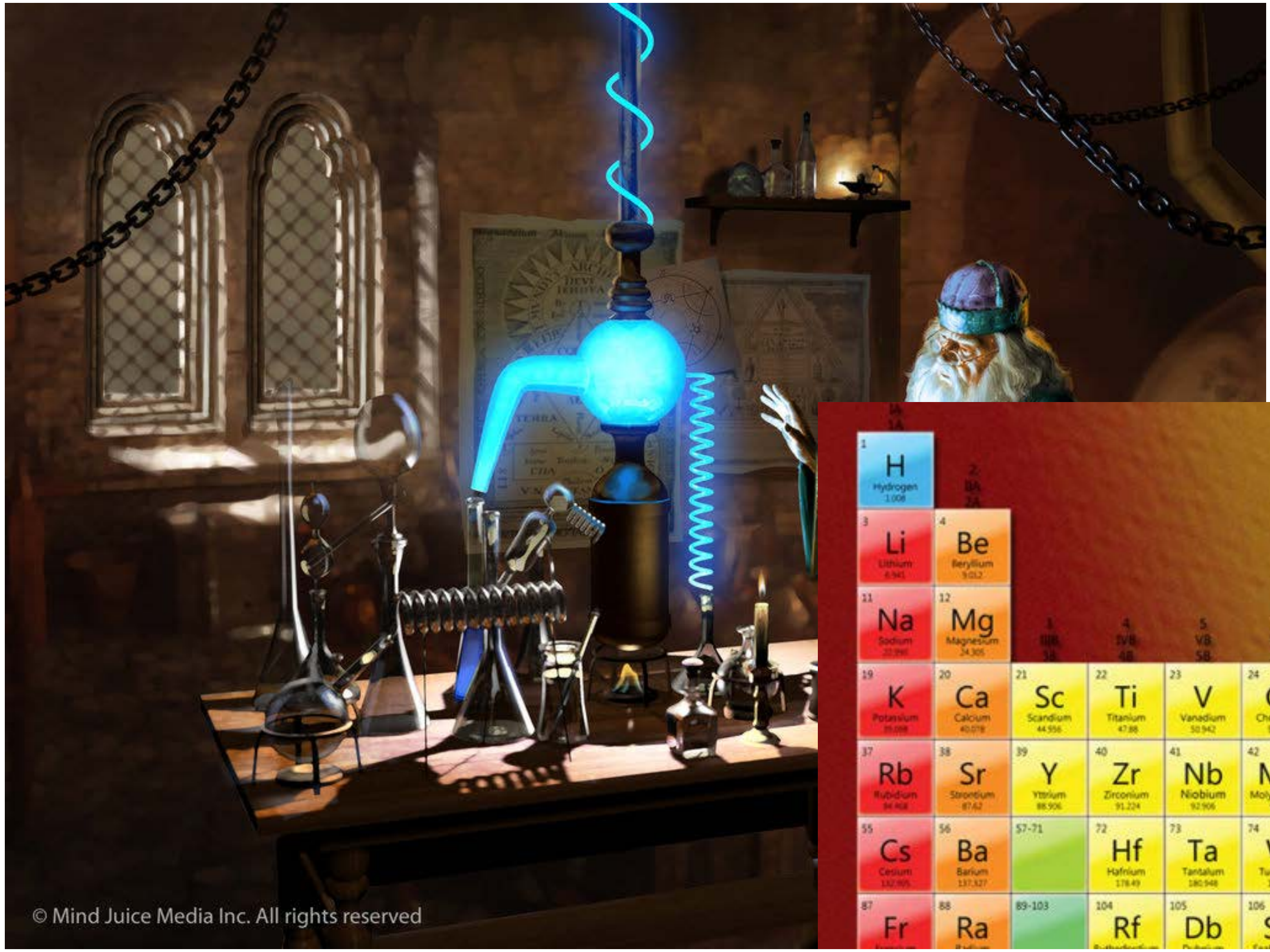
GANs: example results

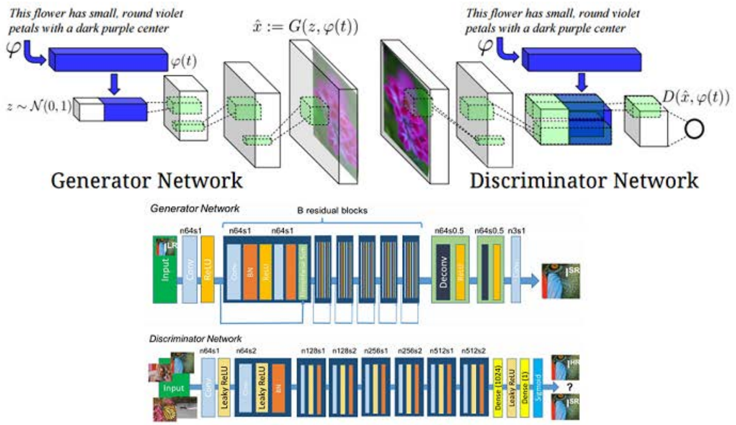
Generated bedrooms [Radford et al., 2016]

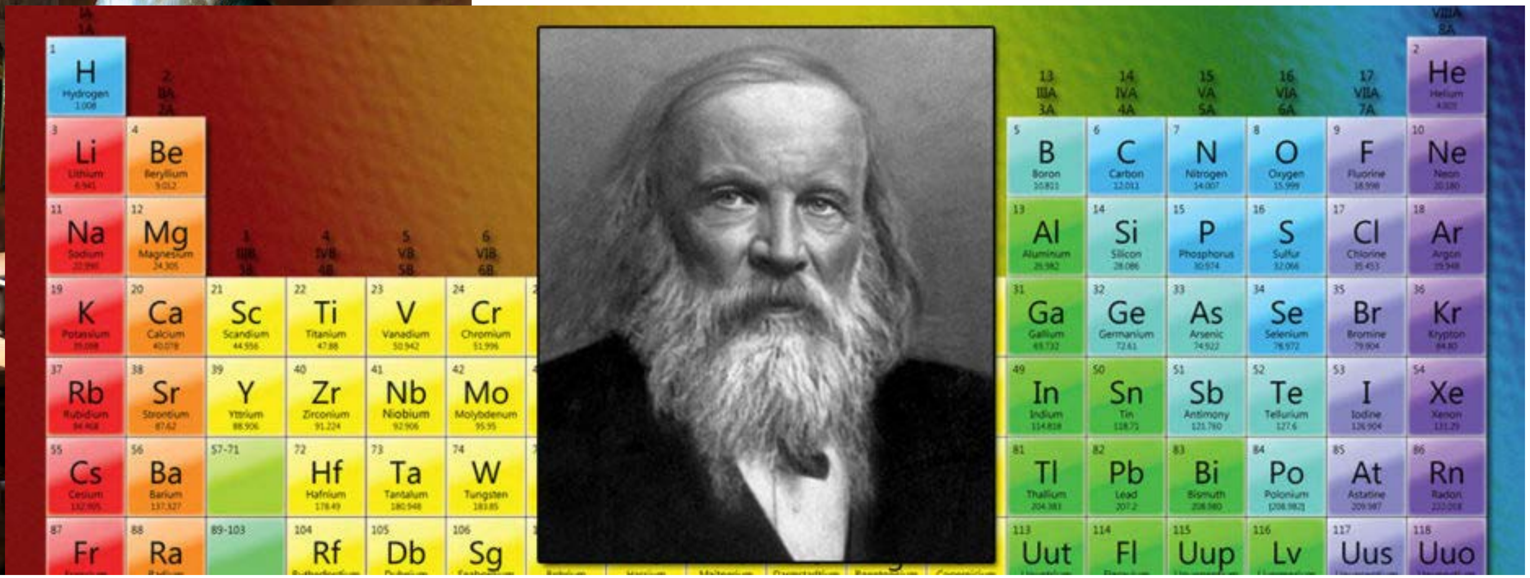
The Zoo of DGMs

- Variational autoencoders (VAEs) [Kingma & Welling, 2014]
  - Adversarial autoencoder [Makhzani et al., 2015]
  - Importance weighted autoencoder [Burda et al., 2015]
  - Implicit variational autoencoder [Mescheder., 2017]
- Generative adversarial networks (GANs) [Goodfellos et al., 2014]
  - InfoGAN [Chen et al., 2016]
  - CycleGAN [Zhu et al., 2017]
  - Wasserstein GAN [Arjovsky et al., 2017]
- Autoregressive neural networks
  - PixelRNN / PixelCNN [Oord et al., 2016]
  - RNN (e.g., for language modeling)
- Generative moment matching networks (GMMNs) [Li et al., 2015; Dziugaite et al., 2015]
- Retricted Boltzmann Machines (RBMs) [Smolensky, 1986]

Alchemy Vs Chemistry







Outline

- Theoretical backgrounds of deep generative models
  - Wake sleep algorithm
  - Variational autoencoders
  - Generative adversarial networks
- A unified view of deep generative models
  - New formulations of deep generative models
  - Symmetric modeling of latent and visible variables

Z Hu, Z YANG, R Salakhutdinov, E Xing,
"**On Unifying Deep Generative Models**", arxiv 1706.00550

Generative Adversarial Nets (GANs):

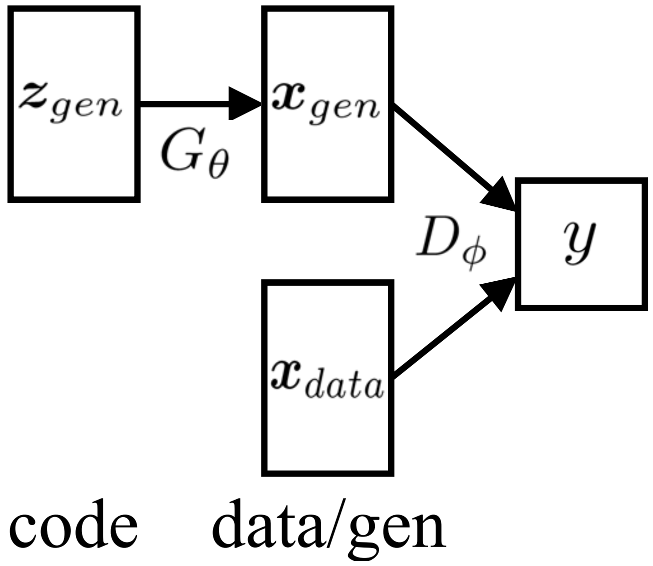
- Implicit distribution over $\boldsymbol{x} \sim p_\theta(\boldsymbol{x}|y)$

$$p_\theta(\boldsymbol{x}|y) = \begin{cases} p_{g_\theta}(\boldsymbol{x}) & y = 0 \\ p_{data}(\boldsymbol{x}) & y = 1. \end{cases}$$

(distribution of generated images)

(distribution of real images)

- $\boldsymbol{x} \sim p_{g_\theta}(\boldsymbol{x}) \Leftrightarrow \boldsymbol{x} = G_\theta(\boldsymbol{z}),\ \boldsymbol{z} \sim p(\boldsymbol{z}|y = 0)$
- $\boldsymbol{x} \sim p_{data}(\boldsymbol{x})$
  - the code space of $\boldsymbol{z}$ is degenerated
  - sample directly from data



A new formulation

- Rewrite GAN objectives in the ”variational-EM” format
- Recap: conventional formulation:

$$\max_{\boldsymbol{\phi}} \mathcal{L}_\phi = \mathbb{E}_{\boldsymbol{x}=G_\theta(\boldsymbol{z}),\boldsymbol{z}\sim p(\boldsymbol{z}|y=0)}\left[\log(1-D_\phi(\boldsymbol{x}))\right] + \mathbb{E}_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}\left[\log D_\phi(\boldsymbol{x})\right]$$

$$\max_{\boldsymbol{\theta}} \mathcal{L}_\theta = \mathbb{E}_{\boldsymbol{x}=G_\theta(\boldsymbol{z}),\boldsymbol{z}\sim p(\boldsymbol{z}|y=0)}\left[\log D_\phi(\boldsymbol{x})\right] + \mathbb{E}_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}\left[\log(1-D_\phi(\boldsymbol{x}))\right]$$

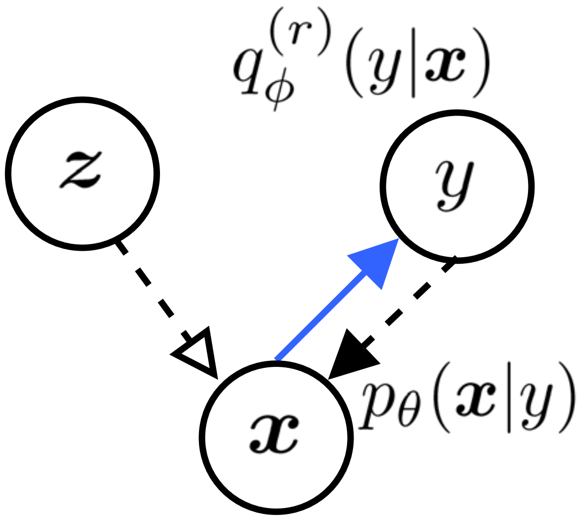
$$= \mathbb{E}_{\boldsymbol{x}=G_\theta(\boldsymbol{z}),\boldsymbol{z}\sim p(\boldsymbol{z}|y=0)}\left[\log D_\phi(\boldsymbol{x})\right]$$

- Rewrite in the new form
  - Implicit distribution over $\boldsymbol{x} \sim p_\theta(\boldsymbol{x}|y)$

$$\boldsymbol{x} = G_\theta(\boldsymbol{z}),\ \boldsymbol{z} \sim p(\boldsymbol{z}|y)$$

  - Discriminator distribution $q_\phi(y|\boldsymbol{x})$

$$q_\phi^r(y|\boldsymbol{x}) = q_\phi(1-y|\boldsymbol{x})$$



$$\max_{\boldsymbol{\phi}} \mathcal{L}_\phi = \mathbb{E}_{p_\theta(\boldsymbol{x}|y)p(y)}\left[\log q_\phi(y|\boldsymbol{x})\right]$$

$$\max_{\boldsymbol{\theta}} \mathcal{L}_\theta = \mathbb{E}_{p_\theta(\boldsymbol{x}|y)p(y)}\left[\log q_\phi^r(y|\boldsymbol{x})\right]$$

GANs vs. Variational EM

Variational EM

- Objectives

$$\max_{\phi} \mathcal{L}_{\boldsymbol{\phi},\boldsymbol{\theta}} = \mathbb{E}_{q_{\phi}(z|x)}[\log p_{\theta}(x|z)] + KL\left(q_{\phi}(z|x)||p(z)\right)$$

$$\max_{\theta} \mathcal{L}_{\boldsymbol{\phi},\boldsymbol{\theta}} = \mathbb{E}_{q_{\phi}(z|x)}[\log p_{\theta}(x|z)] + KL\left(q_{\phi}(z|x)||p(z)\right)$$

  - Single objective for both θ and ϕ
  - Extra prior regularization by $p(z)$
- The reconstruction term: maximize the conditional log-likelihood of x with the generative distribution $p_{\theta}(x|z)$ conditioning on the latent code z inferred by $q_{\phi}(z|x)$

- $p_{\theta}(x|z)$ is the generative model
- $q_{\phi}(z|x)$ is the inference model

GAN

- Objectives

$$\max_{\boldsymbol{\phi}} \mathcal{L}_{\phi} = \mathbb{E}_{p_{\theta}(\boldsymbol{x}|y)p(y)} [\log q_{\phi}(y|\boldsymbol{x})]$$

$$\max_{\boldsymbol{\theta}} \mathcal{L}_{\theta} = \mathbb{E}_{p_{\theta}(\boldsymbol{x}|y)p(y)} \left[\log q^{r}_{\phi}(y|\boldsymbol{x})\right]$$

  - Two objectives
  - Have global optimal state in the game theoretic view
- The objectives: maximize the conditional log-likelihood of y (or $1-y$) with the distribution $q_{\phi}(y|x)$ conditioning on data/generation x inferred by $p_{\theta}(x|y)$

- Interpret $q_{\phi}(y|x)$ as the generative model
- Interpret $p_{\theta}(x|y)$ as the inference model

GANs vs. Variational EM

- Interpret $\boldsymbol{x}$ as latent variables
- Interpret generation of $\boldsymbol{x}$ as performing inference over latent

Variational EM

- Objectives

$$\max_{\phi} \mathcal{L}_{\phi,\theta} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] + KL\left(q_\phi(z|x)||p(z)\right)$$

$$\max_{\theta} \mathcal{L}_{\phi,\theta} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] + KL\left(q_\phi(z|x)||p(z)\right)$$

  - Single objective for both θ and ϕ
  - Extra prior regularization by $p(z)$
- The reconstruction term: maximize the conditional log-likelihood of x with the generative distribution $p_\theta(x|z)$ conditioning on the latent code z inferred by $q_\phi(z|x)$

- $p_\theta(x|z)$ is the generative model
- $q_\phi(z|x)$ is the inference model

GAN

- Objectives

$$\max_{\phi} \mathcal{L}_{\phi} = \mathbb{E}_{p_\theta(\boldsymbol{x}|y)p(y)}\left[\log q_\phi(y|\boldsymbol{x})\right]$$

$$\max_{\theta} \mathcal{L}_{\theta} = \mathbb{E}_{p_\theta(\boldsymbol{x}|y)p(y)}\left[\log q_\phi^r(y|\boldsymbol{x})\right]$$

  - Two objectives
  - Have global optimal state in the game theoretic view
- The objectives: maximize the conditional log-likelihood of y (or $1-y$) with the distribution $q_\phi(y|x)$ conditioning on data/generation x inferred by $p_\theta(x|y)$

- Interpret $q_\phi(y|x)$ as the generative model
- Interpret $p_\theta(x|y)$ as the inference model

GANs: minimizing KLD

- As in Variational EM, we can further rewrite in the form of minimizing KLD to reveal more insights into the optimization problem
- For each optimization step of $p_\theta(\boldsymbol{x}|y)$ at point $(\theta = \theta_0, \phi = \phi_0)$, let
  - $p(y)$: uniform prior distribution
  - $p_{\theta=\theta_0}(\boldsymbol{x}) = \mathrm{E}_{p(y)}[p_{\theta=\theta_0}(\boldsymbol{x}|y)]$
  - $q^r(\boldsymbol{x}|y) \propto q^r_{\phi=\phi_0}(y|\boldsymbol{x}) p_{\theta=\theta_0}(\boldsymbol{x})$
- *Lemma 1*: The updates of $\boldsymbol{\theta}$ at $\boldsymbol{\theta}_0$ have

$$\nabla_\theta \left[-\mathbb{E}_{p_\theta(\boldsymbol{x}|y)p(y)} \left[\log q^r_{\phi=\phi_0}(y|\boldsymbol{x}) \right] \right] \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} =$$

$$\nabla_\theta \left[\mathbb{E}_{p(y)} \left[KL\left(p_\theta(\boldsymbol{x}|y) \| q^r(\boldsymbol{x}|y) \right) \right] - JSD\left(p_\theta(\boldsymbol{x}|y=0) \| p_\theta(\boldsymbol{x}|y=1) \right) \right] \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$$

  - KL: KL divergence
  - JSD: Jensen-shannon divergence

GANs: minimizing KLD

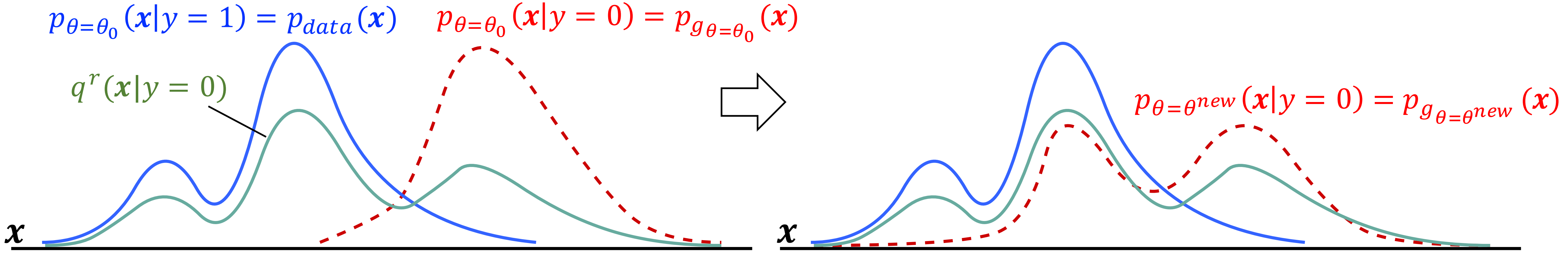
- *Lemma 1*: The updates of $\boldsymbol{\theta}$ at $\boldsymbol{\theta}_0$ have

$$\nabla_\theta \Big[- \mathbb{E}_{p_\theta(\boldsymbol{x}|y)p(y)} \left[\log q^r_{\phi=\phi_0}(y|\boldsymbol{x}) \right] \Big] \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} =$$

$$\nabla_\theta \Big[\mathbb{E}_{p(y)} \left[\mathrm{KL}\left(p_\theta(\boldsymbol{x}|y) \| q^r(\boldsymbol{x}|y) \right) \right] - \mathrm{JSD}\left(p_\theta(\boldsymbol{x}|y=0) \| p_\theta(\boldsymbol{x}|y=1) \right) \Big] \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$$

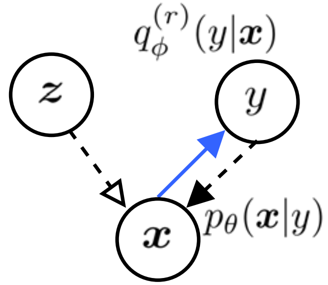
- Connection to variational inference
  - See $\boldsymbol{x}$ as latent variables, y as visible
  - $p_{\theta=\theta_0}(\boldsymbol{x})$: prior distribution
  - $q^r(\boldsymbol{x}|y) \propto q^r_{\phi=\phi_0}(y|\boldsymbol{x}) p_{\theta=\theta_0}(\boldsymbol{x})$: posterior distribution
  - $p_\theta(\boldsymbol{x}|y)$: variational distribution
    - Amortized inference: updates model parameter $\boldsymbol{\theta}$
- Suggests relations to VAEs, as we will explore shortly

GANs: minimizing KLD

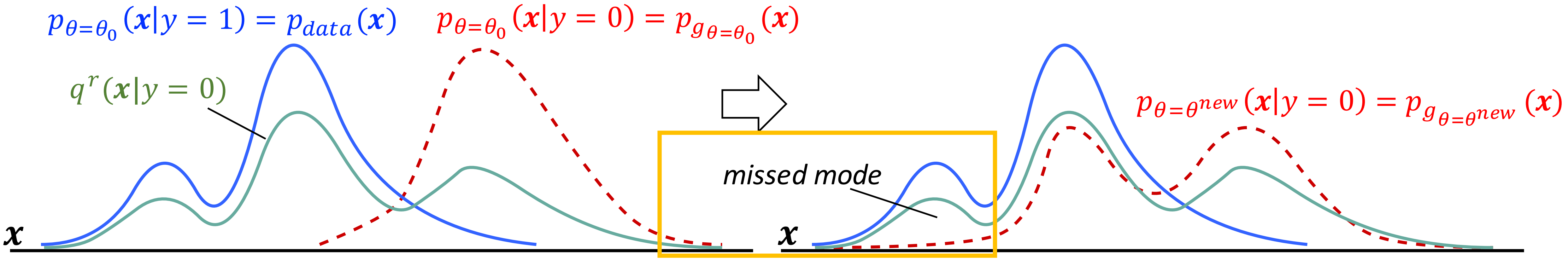


- Minimizing the KLD drives $p_{g_\theta}(\boldsymbol{x})$ to $p_{data}(\boldsymbol{x})$
  - By definition: $p_{\theta=\theta_0}(\boldsymbol{x}) = \mathrm{E}_{p(y)}[p_{\theta=\theta_0}(\boldsymbol{x}|y)] = \left(p_{g_{\theta=\theta_0}}(\boldsymbol{x}) + p_{data}(\boldsymbol{x})\right) / 2$
  - $\mathrm{KL}\left(p_\theta(x|y=1)||q^r(x|y=1)\right) = \mathrm{KL}\left(p_{data}(x)||q^r(x|y=1)\right)$: constant, no free parameters
  - $\mathrm{KL}\left(p_\theta(x|y=0)||q^r(x|y=0)\right) = \mathrm{KL}\left(p_{g_\theta}(x)||q^r(x|y=0)\right)$: parameter θ to optimize
    - $q^r(\boldsymbol{x}|y=0) \propto q^r_{\phi=\phi_0}(y=0|\boldsymbol{x}) p_{\theta=\theta_0}(\boldsymbol{x})$
      - seen as a mixture of $p_{g_{\theta=\theta_0}}(\boldsymbol{x})$ and $p_{data}(\boldsymbol{x})$
      - mixing weights induced from $q^r_{\phi=\phi_0}(y=0|\boldsymbol{x})$
    - Drives $p_{g_\theta}(\boldsymbol{x}|y)$ to mixture of $p_{g_{\theta=\theta_0}}(\boldsymbol{x})$ and $p_{data}(\boldsymbol{x})$

      $\Rightarrow$ Drives $p_{g_\theta}(\boldsymbol{x})$ to $p_{data}(\boldsymbol{x})$



GANs: minimizing KLD



- Missing mode phenomena of GANs
  - Asymmetry of KLD
    - Concentrates $p_\theta(\boldsymbol{x}|y=0)$ to large modes of $q^r(\boldsymbol{x}|y)$

      $\Rightarrow p_{g_\theta}(\boldsymbol{x})$ misses modes of $p_{data}(\boldsymbol{x})$
  - Symmetry of JSD
    - Does not affect the behavior of mode missing

$$\mathrm{KL}\left(p_{g_\theta}(x)||q^r(x|y=0)\right) = \int p_{g_\theta}(x)\log\frac{p_{g_\theta}(x)}{q^r(x|y=0)}dx$$

- Large positive contribution to the KLD in the regions of x space where $q^r(x|y=0)$ is small, unless $p_{g_\theta}(x)$ is also small
- $\Rightarrow p_{g_\theta}(x)$ tends to avoid regions where $q^r(x|y=0)$ is small

GANs: minimizing KLD

- *Lemma 1*: The updates of $\boldsymbol{\theta}$ at $\boldsymbol{\theta}_0$ have

$$\nabla_\theta \left[-\mathbb{E}_{p_\theta(\boldsymbol{x}|y)p(y)} \left[\log q^r_{\phi_0}(y|\boldsymbol{x}) \right] \right] \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} =$$

$$\nabla_\theta \left[\mathbb{E}_{p(y)} \left[KL\left(p_\theta(\boldsymbol{x}|y) \| q^r(\boldsymbol{x}|y) \right) \right] - JSD\left(p_\theta(\boldsymbol{x}|y=0) \| p_\theta(\boldsymbol{x}|y=1) \right) \right] \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$$

- No assumption on optimal discriminator $q^r_{\phi_0}(y|\boldsymbol{x})$
  - Previous results usually rely on (near) optimal discriminator
    - $q^*(y=1|\boldsymbol{x}) = p_{data}(\boldsymbol{x})/(p_{data}(\boldsymbol{x}) + p_g(\boldsymbol{x}))$
  - Optimality assumption is impractical: limited expressiveness of D_ϕ [Arora et al 2017]
  - Our result is a generalization of the previous theorem [Arjovsky & Bottou 2017]
    - Plug the optimal discriminator into the above equation, we recover the theorem

$$\nabla_\theta \left[-\mathbb{E}_{p_\theta(\boldsymbol{x}|y)p(y)} \left[\log q^r_{\phi_0}(y|\boldsymbol{x}) \right] \right] \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \nabla_\theta \left[\frac{1}{2} \mathrm{KL}\left(p_{g_\theta} \| p_{data} \right) - \mathrm{JSD}\left(p_{g_\theta} \| p_{data} \right) \right] \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$$

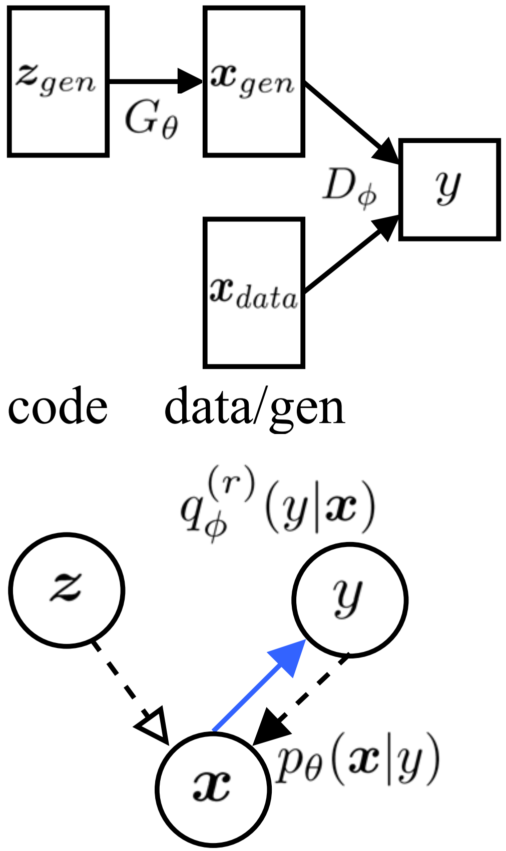
    - Give insights on the generator training when discriminator is optimal

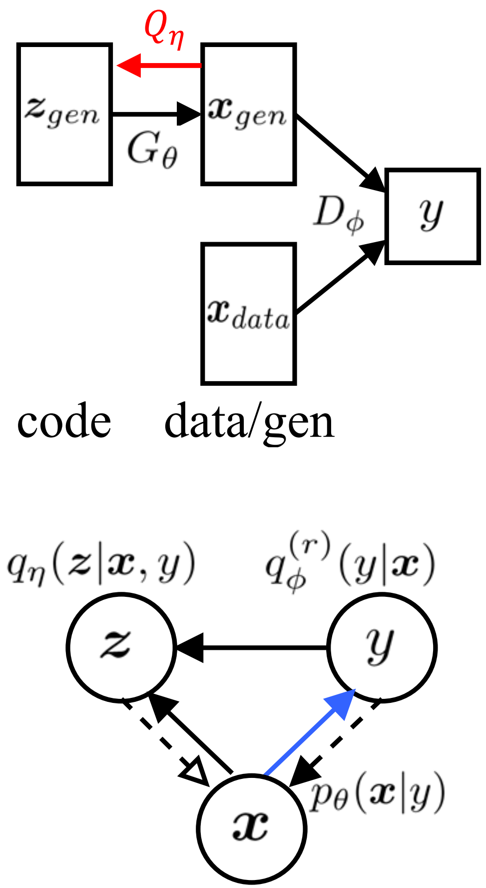
GANs: minimizing KLD

In summary:

- Reveal connection to variational inference
  - Build connections to VAEs (slides soon)
  - Inspire new model variants based on the connections
- Offer insights into the generator training
  - Formal explanation of the missing mode behavior of GANs
  - Still hold when the discriminator does not achieve its optimum at each iteration

GANs vs InfoGAN





$$\max_{\boldsymbol{\phi}} \mathcal{L}_\phi = \mathbb{E}_{p_\theta(\boldsymbol{x}|y)p(y)}\left[\log q_\phi(y|\boldsymbol{x})\right]$$

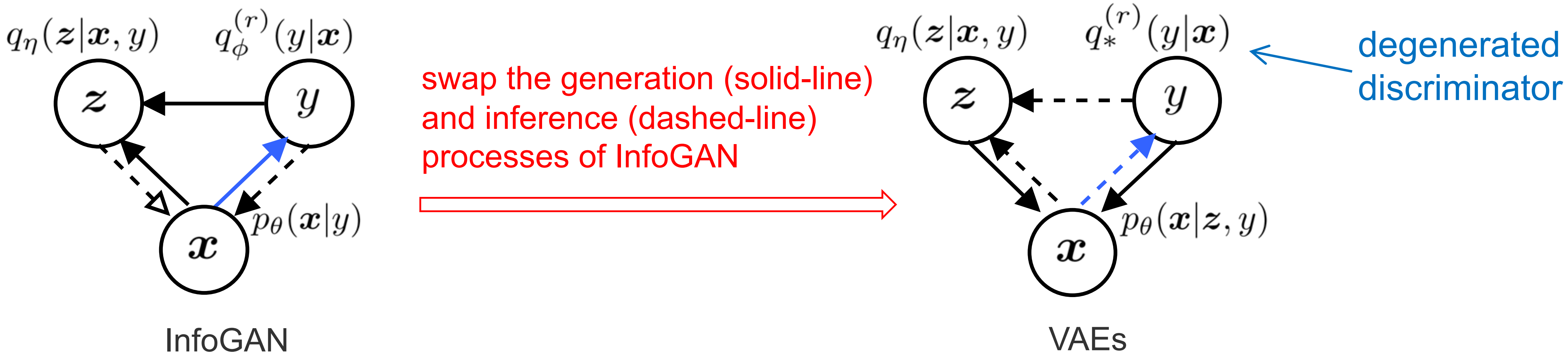
$$\max_{\boldsymbol{\theta}} \mathcal{L}_\theta = \mathbb{E}_{p_\theta(\boldsymbol{x}|y)p(y)}\left[\log q_\phi^r(y|\boldsymbol{x})\right]$$

$$\max_{\boldsymbol{\phi}} \mathcal{L}_\phi = \mathbb{E}_{p_\theta(\boldsymbol{x}|y)p(y)}\left[\log q_\eta(\boldsymbol{z}|\boldsymbol{x}, y) q_\phi(y|\boldsymbol{x})\right]$$

$$\max_{\boldsymbol{\theta},\boldsymbol{\eta}} \mathcal{L}_{\theta,\eta} = \mathbb{E}_{p_\theta(\boldsymbol{x}|y)p(y)}\left[\log q_\eta(\boldsymbol{z}|\boldsymbol{x}, y) q_\phi^r(y|\boldsymbol{x})\right]$$

Relates VAEs with GANs

- Resemblance of GAN generator learning to variational inference
  - Suggest strong relations between VAEs and GANs
- Indeed, VAEs are basically minimizing KLD with an opposite direction, and with a degenerated adversarial discriminator



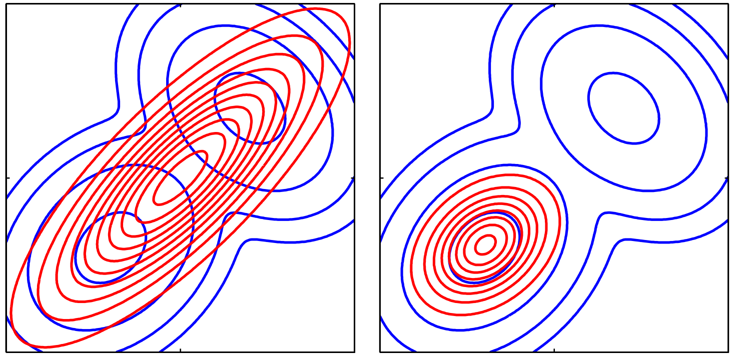
GANs vs VAEs side by side

	GANs (InfoGAN)	VAEs
Generative distribution	$p_\theta(\boldsymbol{x}\|y) = \begin{cases} p_{g_\theta}(\boldsymbol{x}) & y=0 \\ p_{data}(\boldsymbol{x}) & y=1. \end{cases}$	$p_\theta(\boldsymbol{x}\|\boldsymbol{z}, y) = \begin{cases} p_\theta(\boldsymbol{x}\|\boldsymbol{z}) & y=0 \\ p_{data}(\boldsymbol{x}) & y=1. \end{cases}$
Discriminator distribution	$q_\phi(y\|\boldsymbol{x})$	$q_*(y\|\boldsymbol{x})$, perfect, degenerated
$\boldsymbol{z}$-inference model	$q_\eta(\boldsymbol{z}\|\boldsymbol{x}, y)$ of InfoGAN	$q_\eta(\boldsymbol{z}\|\boldsymbol{x}, y)$
KLD to minimize	$\min_\theta \mathrm{KL}\,(p_\theta(\boldsymbol{x}\|y) \parallel q^r(\boldsymbol{x}\|\boldsymbol{z}, y))$ $\sim \min_\theta \mathrm{KL}(P_\theta \parallel Q)$	$\min_\theta \mathrm{KL}\left(q_\eta(\boldsymbol{z}\|\boldsymbol{x}, y) q_*^r(y\|\boldsymbol{x}) \parallel p_\theta(\boldsymbol{z}, y\|\boldsymbol{x})\right)$ $\sim \min_\theta \mathrm{KL}(Q \parallel P_\theta)$

GANs vs VAEs side by side

	GANs (InfoGAN)	VAEs
KLD to minimize	$\min_\theta \mathrm{KL}\left(p_\theta(\boldsymbol{x}\vert y) \,\vert\vert\, q^r(\boldsymbol{x}\vert\boldsymbol{z}, y)\right)$ $\sim \min_\theta \mathrm{KL}(P_\theta \,\vert\vert\, Q)$	$\min_\theta \mathrm{KL}(q_\eta(\boldsymbol{z}\vert\boldsymbol{x}, y) q_*^r(y\vert\boldsymbol{x}) \,\vert\vert\, p_\theta(\boldsymbol{z}, y\vert\boldsymbol{x}))$ $\sim \min_\theta \mathrm{KL}(Q \,\vert\vert\, P_\theta)$

- Asymmetry of KLDs inspires combination of GANs and VAEs
  - GANs: $\min_\theta \mathrm{KL}(P_\theta || Q)$ tends to missing mode
  - VAEs: $\min_\theta \mathrm{KL}(Q || P_\theta)$ tends to cover regions with small values of p_{data}



Mode covering Mode missing

[Figure courtesy: PRML]

Mutual exchanges of ideas: augment the loss

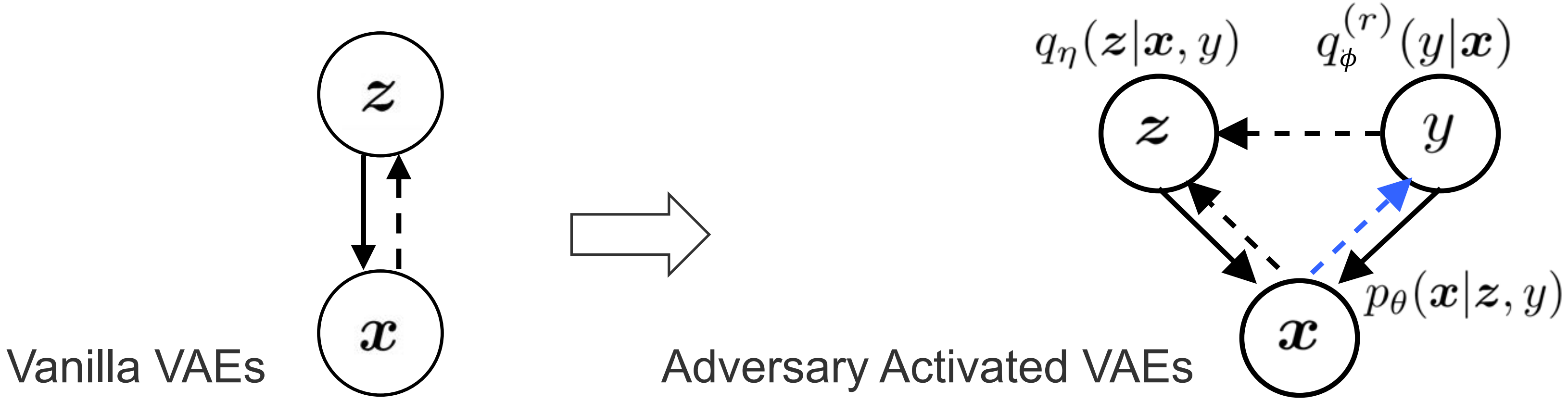
	GANs (InfoGAN)	VAEs
KLD to minimize	$\min_\theta \text{KL}\left(p_\theta(\boldsymbol{x}\|y) \,\|\|\, q^r(\boldsymbol{x}\|\boldsymbol{z},y)\right)$ $\sim \min_\theta \text{KL}(P_\theta \,\|\|\, Q)$	$\min_\theta \text{KL}(q_\eta(\boldsymbol{z}\|\boldsymbol{x},y)q_*^r(y\|\boldsymbol{x}) \,\|\|\, p_\theta(\boldsymbol{z},y\|\boldsymbol{x}))$ $\sim \min_\theta \text{KL}(Q \,\|\|\, P_\theta)$

- Asymmetry of KLDs inspires combination of GANs and VAEs
  - GANs: $\min_\theta \text{KL}(P_\theta\|\|Q)$ tends to missing mode
  - VAEs: $\min_\theta \text{KL}(Q\|\|P_\theta)$ tends to cover regions with small values of p_{data}
  - Augment VAEs with GAN loss [Larsen et al., 2016]
    - Alleviate the mode covering issue of VAEs
    - Improve the sharpness of VAE generated images
  - Augment GANs with VAE loss [Che et al., 2017]
    - Alleviate the mode missing issue of GAN

Mutual exchanges of ideas: augment the model

	GANs (InfoGAN)	VAEs
Discriminator distribution	$q_\phi(y\|\boldsymbol{x})$	$q_*(y\|\boldsymbol{x})$, perfect, degenerated

- Activate the adversarial mechanism in VAEs
  - Enable adaptive incorporation of fake samples for learning
  - Straightforward derivation by making symbolic analog to GANs



Adversary Activated VAEs (AAVAE)

- Vanilla VAEs:

$$\max_{\boldsymbol{\theta},\boldsymbol{\eta}} \mathcal{L}_{\theta,\eta}^{\text{vae}} = \mathbb{E}_{p_{\theta_0}(\boldsymbol{x})} \left[\mathbb{E}_{q_\eta(\boldsymbol{z}|\boldsymbol{x},y) q_*^r(y|\boldsymbol{x})} \left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z},y) \right] - \mathrm{KL}(q_\eta(\boldsymbol{z}|\boldsymbol{x},y) q_*^r(y|\boldsymbol{x}) \| p(\boldsymbol{z}|y)p(y)) \right]$$

- Replace $q_*(y|\boldsymbol{x})$ with learnable one $q_\phi(y|\boldsymbol{x})$ with parameters $\boldsymbol{\phi}$
  - As usual, denote reversed distribution $q_\phi^r(y|x) = q_\phi(y|x)$

$$\max_{\boldsymbol{\theta},\boldsymbol{\eta}} \mathcal{L}_{\theta,\eta}^{\text{aavae}} = \mathbb{E}_{p_{\theta_0}(\boldsymbol{x})} \left[\mathbb{E}_{q_\eta(\boldsymbol{z}|\boldsymbol{x},y) q_\phi^r(y|\boldsymbol{x})} \left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z},y) \right] - \mathrm{KL}(q_\eta(\boldsymbol{z}|\boldsymbol{x},y) q_\phi^r(y|\boldsymbol{x}) \| p(\boldsymbol{z}|y)p(y)) \right]$$

AAVAE: empirical results

- Applied the adversary activating method on
  - vanilla VAEs
  - class-conditional VAEs (CVAE)
  - semi-supervised VAEs (SVAE)
- Evaluated test-set variational lower bound on MNIST
  - The higher the better

Train Data Size	VAE	AA-VAE	CVAE	AA-CVAE	SVAE	AA-SVAE
1%	-122.89	**-122.15**	-125.44	**-122.88**	-108.22	**-107.61**
10%	-104.49	**-103.05**	-102.63	**-101.63**	-99.44	**-98.81**
100%	-92.53	**-92.42**	-93.16	**-92.75**	—	—

- X-axis: the ratio of training data for learning: (1%, 10%, 100%)
- Y-axis: value of test-set lower bound

AAVAE: empirical results

- Evaluated classification accuracy of SVAE and AA-SVAE

	1%	10%
SVAE	0.9412±.0039	0.9768±.0009
AASVAE	**0.9425±.0045**	**0.9797±.0010**

  - Used 1% and 10% data labels in MNIST

Importance weighted GANs (IWGAN)

- Generator learning in vanilla GANs

$$\max_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{x} \sim p_\theta(\boldsymbol{x}|y)p(y)} \left[\log q^r_{\phi_0}(y|\boldsymbol{x}) \right]$$

- Generator learning in IWGAN

$$\max_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \sim p_\theta(\boldsymbol{x}|y)p(y)} \left[\sum_{i=1}^{k} \frac{q^r_{\phi_0}(y|\boldsymbol{x}_i)}{q_{\phi_0}(y|\boldsymbol{x}_i)} \log q^r_{\phi_0}(y|\boldsymbol{x}_i) \right]$$

  - Assigns higher weights to samples that are more realistic and fool the discriminator better

IWGAN: empirical results

- Evaluated on MNIST and SVHN
- Used pretrained NN to evaluate:
  - Inception scores of samples from GANs and IW-GAN
    - Confidence of a pre-trained classifier on generated samples + diversity of generated samples

	MNIST	SVHN
GAN	8.34±.03	5.18±.03
IWGAN	**8.45±.04**	**5.34±.03**

  - Classification accuracy of samples from CGAN and IW-CGAN

	MNIST	SVHN
CGAN	0.985±.002	0.797±.005
IWCGAN	**0.987±.002**	**0.798±.006**

Symmetric modeling of latent & visible variables

Empirical distributions over visible variables

- Impossible to be explicit distribution
  - The only information we have is the observe data examples
  - Do not know the true parametric form of data distribution
- Naturally an implicit distribution
  - Easy to sample from, hard to evaluate likelihood

Prior distributions over latent variables

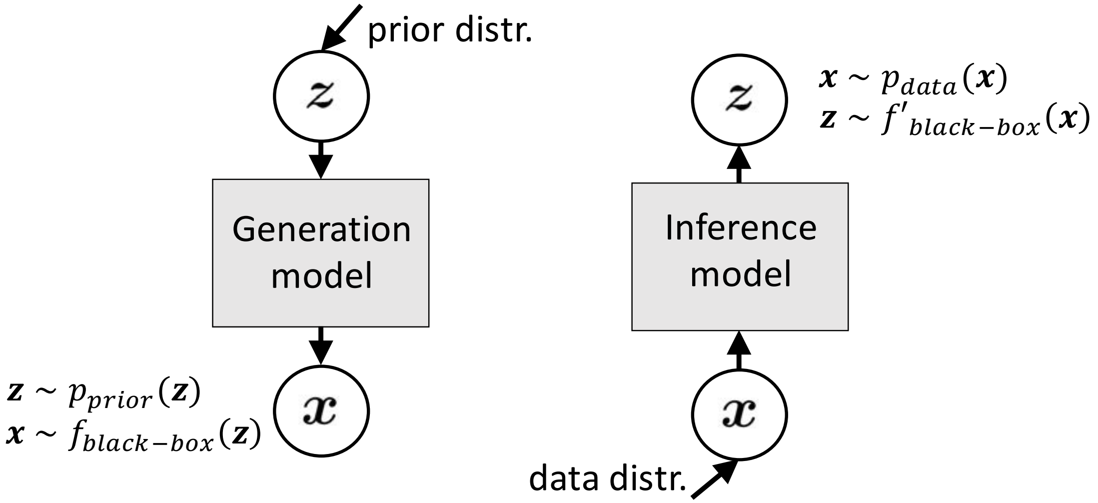
- Traditionally defined as explicit distributions, e.g., Gaussian prior distribution
  - Amiable for likelihood evaluation
  - We can assume the parametric form according to our prior knowledge
- New tools to allow implicit priors and models
  - GANs, density ratio estimation, approximate Bayesian computations
  - E.g., adversarial autoencoder [Makhzani et al., 2015] replaces the Gaussian prior of vanilla VAEs with implicit priors

Symmetric modeling of latent & visible variables

- No difference in terms of formulations
  - with implicit distributions and black-box NN models
  - just swap the symbols x and z

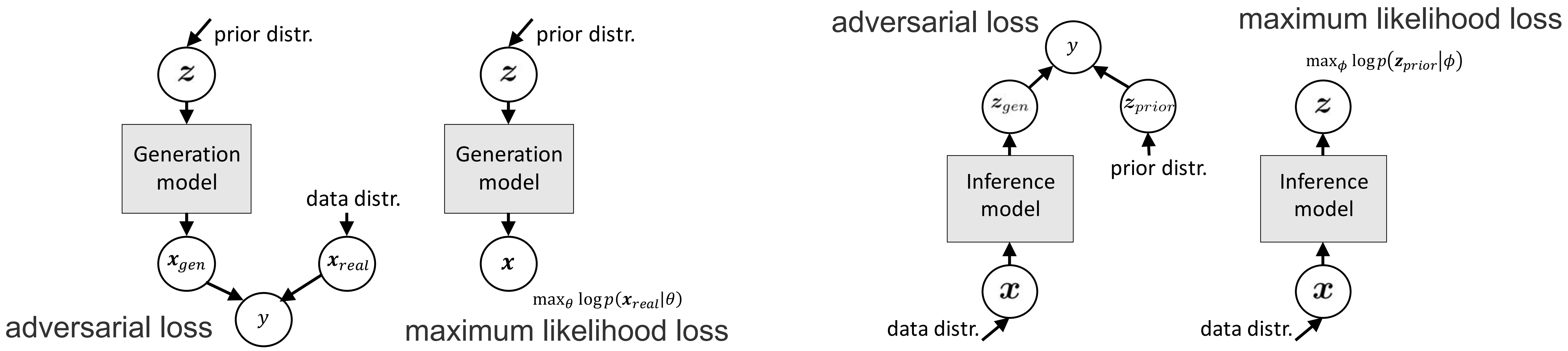
$$\boldsymbol{z} \sim p_{prior}(\boldsymbol{z})$$
$$\boldsymbol{x} \sim f_{black-box}(\boldsymbol{z})$$

$$\boldsymbol{x} \sim p_{data}(\boldsymbol{x})$$
$$\boldsymbol{z} \sim f'_{black-box}(\boldsymbol{x})$$



Symmetric modeling of latent & visible variables

- No difference in terms of formulations
  - with implicit distributions and black-box NN models
- Difference in terms of space complexity
  - depend on the problem at hand
  - choose appropriate tools:
    - implicit/explicit distribution, adversarial/maximum-likelihood optimization, …



Conclusions

Z Hu, Z YANG, R Salakhutdinov, E Xing,
"**On Unifying Deep Generative Models**", arxiv 1706.00550

- Deep generative models research have a long history
  - Deep blief nets / Helmholtz machines / Predictability Minimization / …
- Unification of deep generative models
  - GANs and VAEs are essentially minimizing KLD in opposite directions
    - Extends two phases of classic wake sleep algorithm, respectively
  - A general formulation framework useful for
    - Analyzing broad class of existing DGM and variants: ADA/InfoGAN/Joint-models/…
    - Inspiring new models and algorithms by borrowing ideas across research fields
- Symmetric view of latent/visible variables
  - No difference in formulation with implicit prior distributions and black-box NN transformations
  - Difference in space complexity: choose appropriate tools

Thank You