Improving Mutual Information Estimation using Annealed and Energy-Based Bounds

Rob Brekelmans*[1]

Greg Ver Steeg [1]

Sicong Huang*[2,3] , Marzyeh Ghassemi [3,4]

Roger Grosse [2,3], Alireza Makhzani [2,3]

1 USC Viterbi School of Engineering Information Sciences Institute

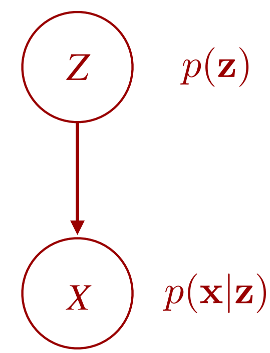
2 University of Toronto

3 Vector Institute

4

Mutual Information Estimation

- General measure of dependence between two RVs X,Z



$$I_p(X;Z) = \mathcal{H}(X) - \mathcal{H}(X|Z) \;\; = \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\big[\log p(\mathbf{x}|\mathbf{z})\big] - \mathbb{E}_{p(\mathbf{x})}\big[\log p(\mathbf{x})\big]$$

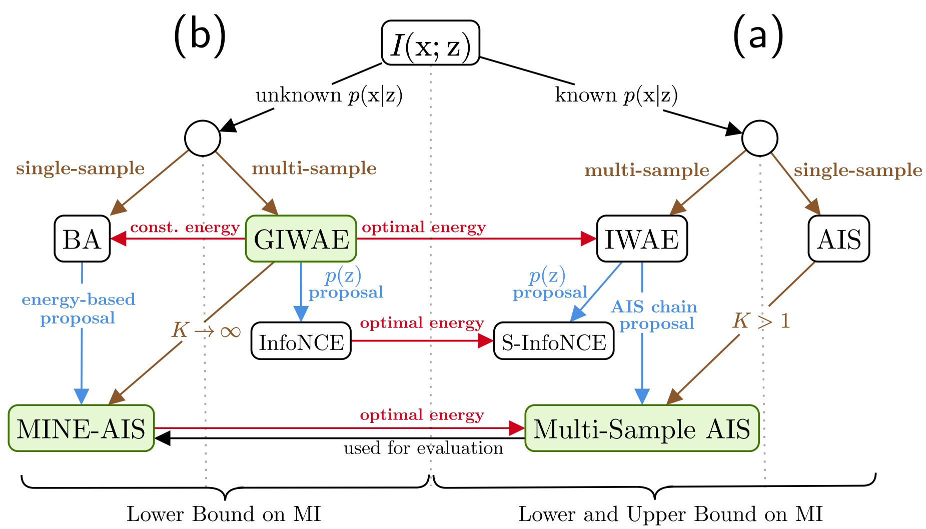
- Estimation from joint samples (x, z) only
  - Nearest neighbors (Kraskov et. al 2004)
  - InfoNCE, MINE (review: Poole et. al 2019)

- Exponential sample complexity!
  - (Gao et. al 2015)
  - (McAllester & Stratos 2020)

Our Contributions

(a) Known $p(\mathbf{x}|\mathbf{z}), p(\mathbf{z})$ Multi-Sample Annealed Importance Sampling

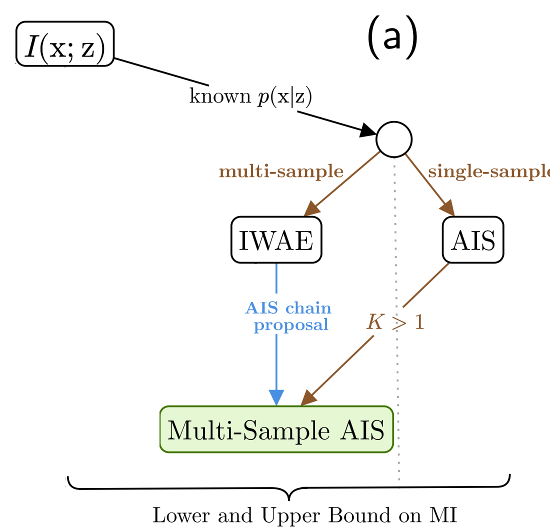
(b) Known $p(\mathbf{z})$ only MINE-AIS, Generalized IWAE



Known Joint Distribution

(a) Known $p(\mathbf{x}|\mathbf{z})$, $p(\mathbf{z})$ Multi-Sample Annealed Importance Sampling

$$I_p(X;Z) = \underbrace{\mathbb{E}_{p(\mathbf{x},\mathbf{z})}\big[\log p(\mathbf{x}|\mathbf{z})\big]}_{\text{unbiased, low variance}} - \mathbb{E}_{p(\mathbf{x})}\big[\underbrace{\log p(\mathbf{x})}_{\text{intractable log partition function}}\big]$$



Extended State Space Approach

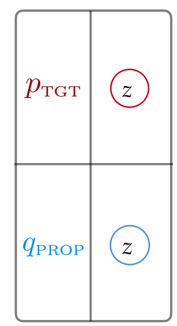
- Construct $p_{\text{TGT}}(\mathbf{x}, \mathbf{z}_{\text{ext}})$, $q_{\text{PROP}}(\mathbf{z}_{\text{ext}}|\mathbf{x})$ such that ratio of normalization constants is $p(\mathbf{x})$

$$\underbrace{\mathbb{E}_{q_{\text{PROP}}(\mathbf{z}_{\text{ext}}|\mathbf{x})}\left[\log \frac{p_{\text{TGT}}(\mathbf{x}, \mathbf{z}_{\text{ext}})}{q_{\text{PROP}}(\mathbf{z}_{\text{ext}}|\mathbf{x})}\right]}_{\text{ELBO}(\mathbf{x})} \leq \log p(\mathbf{x}) \leq \underbrace{\mathbb{E}_{p_{\text{TGT}}(\mathbf{z}_{\text{ext}}|\mathbf{x})}\left[\log \frac{p_{\text{TGT}}(\mathbf{x}, \mathbf{z}_{\text{ext}})}{q_{\text{PROP}}(\mathbf{z}_{\text{ext}}|\mathbf{x})}\right]}_{\text{EUBO}(\mathbf{x})}$$

Standard ELBO and ‘EUBO’

$$p_{\text{TGT}}(\mathbf{x}, \mathbf{z}_{\text{ext}}) = p(\mathbf{x}, \mathbf{z})$$

$$q_{\text{PROP}}(\mathbf{z}_{\text{ext}}|\mathbf{x}) = q_\theta(\mathbf{z}|\mathbf{x})$$



Extended State Space Approach

- Construct $p_{\text{TGT}}(\mathbf{x}, \mathbf{z}_{\text{ext}})$, $q_{\text{PROP}}(\mathbf{z}_{\text{ext}}|\mathbf{x})$ such that ratio of normalization constants is $p(\mathbf{x})$

$$\underbrace{\mathbb{E}_{q_{\text{PROP}}(\mathbf{z}_{\text{ext}}|\mathbf{x})}\left[\log \frac{p_{\text{TGT}}(\mathbf{x}, \mathbf{z}_{\text{ext}})}{q_{\text{PROP}}(\mathbf{z}_{\text{ext}}|\mathbf{x})}\right]}_{\text{ELBO}(\mathbf{x})} \leq \log p(\mathbf{x}) \leq \underbrace{\mathbb{E}_{p_{\text{TGT}}(\mathbf{z}_{\text{ext}}|\mathbf{x})}\left[\log \frac{p_{\text{TGT}}(\mathbf{x}, \mathbf{z}_{\text{ext}})}{q_{\text{PROP}}(\mathbf{z}_{\text{ext}}|\mathbf{x})}\right]}_{\text{EUBO}(\mathbf{x})}$$

Lower Bound Gap

$$D_{\text{KL}}\left[q_{\text{PROP}}(\mathbf{z}_{\text{ext}}|\mathbf{x}) \| p_{\text{TGT}}(\mathbf{z}_{\text{ext}}|\mathbf{x})\right]$$

Upper Bound Gap

$$D_{\text{KL}}\left[p_{\text{TGT}}(\mathbf{z}_{\text{ext}}|\mathbf{x}) \| q_{\text{PROP}}(\mathbf{z}_{\text{ext}}|\mathbf{x})\right]$$

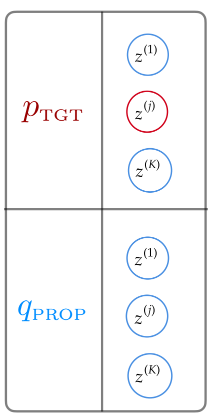
Extended State Space Approach

- Construct $p_{\text{TGT}}(\mathbf{x}, \mathbf{z}_{\text{ext}})$, $q_{\text{PROP}}(\mathbf{z}_{\text{ext}}|\mathbf{x})$ such that ratio of normalization constants is $p(\mathbf{x})$

$$\underbrace{\mathbb{E}_{q_{\text{PROP}}(\mathbf{z}_{\text{ext}}|\mathbf{x})}\left[\log \frac{p_{\text{TGT}}(\mathbf{x}, \mathbf{z}_{\text{ext}})}{q_{\text{PROP}}(\mathbf{z}_{\text{ext}}|\mathbf{x})}\right]}_{\text{ELBO}(\mathbf{x})} \leq \log p(\mathbf{x}) \leq \underbrace{\mathbb{E}_{p_{\text{TGT}}(\mathbf{z}_{\text{ext}}|\mathbf{x})}\left[\log \frac{p_{\text{TGT}}(\mathbf{x}, \mathbf{z}_{\text{ext}})}{q_{\text{PROP}}(\mathbf{z}_{\text{ext}}|\mathbf{x})}\right]}_{\text{EUBO}(\mathbf{x})}$$

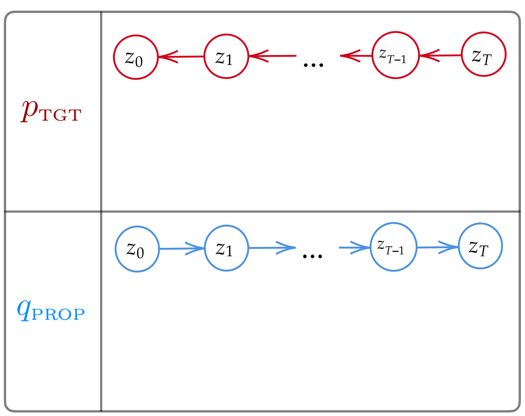
Importance Weighted Autoencoder

(Burda et. al 2016, Sobolev & Vetrov 2019)



Annealed Importance Sampling

(Neal 2001, Grosse et al 2015)



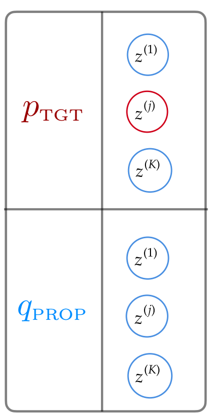
Multi-Sample AIS

- Construct $p_{\text{TGT}}(\mathbf{x}, \mathbf{z}_{\text{ext}})$, $q_{\text{PROP}}(\mathbf{z}_{\text{ext}}|\mathbf{x})$ such that ratio of normalization constants is $p(\mathbf{x})$

$$\underbrace{\mathbb{E}_{q_{\text{PROP}}(\mathbf{z}_{\text{ext}}|\mathbf{x})}\left[\log \frac{p_{\text{TGT}}(\mathbf{x}, \mathbf{z}_{\text{ext}})}{q_{\text{PROP}}(\mathbf{z}_{\text{ext}}|\mathbf{x})}\right]}_{\text{ELBO}(\mathbf{x})} \leq \log p(\mathbf{x}) \leq \underbrace{\mathbb{E}_{p_{\text{TGT}}(\mathbf{z}_{\text{ext}}|\mathbf{x})}\left[\log \frac{p_{\text{TGT}}(\mathbf{x}, \mathbf{z}_{\text{ext}})}{q_{\text{PROP}}(\mathbf{z}_{\text{ext}}|\mathbf{x})}\right]}_{\text{EUBO}(\mathbf{x})}$$

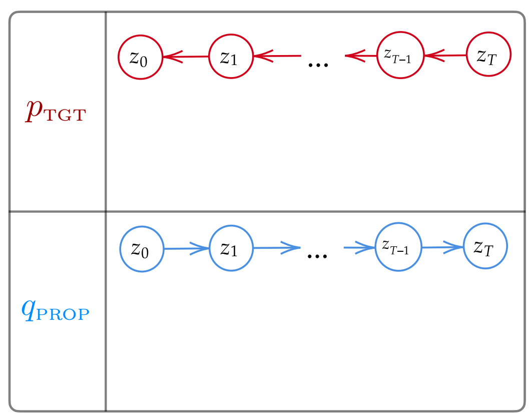
Importance Weighted Autoencoder

(Burda et. al 2016, Sobolev & Vetrov 2019)



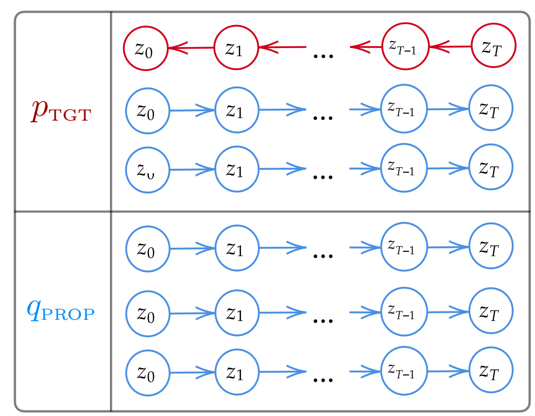
Annealed Importance Sampling

(Neal 2001, Grosse et al 2015)



Multi-Sample AIS

(ours)



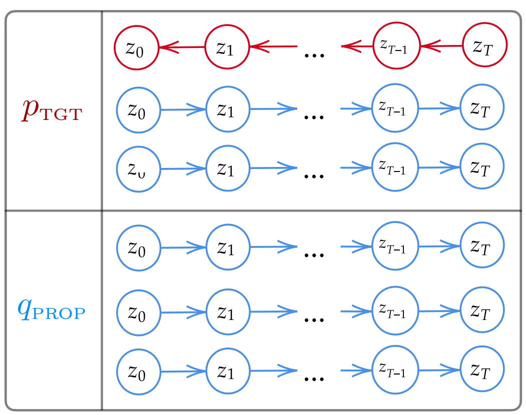
Bidirectional Monte Carlo (BDMC)

- Construct $p_{\text{TGT}}(\mathbf{x}, \mathbf{z}_{\text{ext}})$, $q_{\text{PROP}}(\mathbf{z}_{\text{ext}} | \mathbf{x})$ such that ratio of normalization constants is $p(\mathbf{x})$

$$\underbrace{\mathbb{E}_{q_{\text{PROP}}(\mathbf{z}_{\text{ext}}|\mathbf{x})}\left[\log \frac{p_{\text{TGT}}(\mathbf{x}, \mathbf{z}_{\text{ext}})}{q_{\text{PROP}}(\mathbf{z}_{\text{ext}}|\mathbf{x})}\right]}_{\text{ELBO}(\mathbf{x})} \leq \log p(\mathbf{x}) \leq \underbrace{\mathbb{E}_{p_{\text{TGT}}(\mathbf{z}_{\text{ext}}|\mathbf{x})}\left[\log \frac{p_{\text{TGT}}(\mathbf{x}, \mathbf{z}_{\text{ext}})}{q_{\text{PROP}}(\mathbf{z}_{\text{ext}}|\mathbf{x})}\right]}_{\text{EUBO}(\mathbf{x})}$$

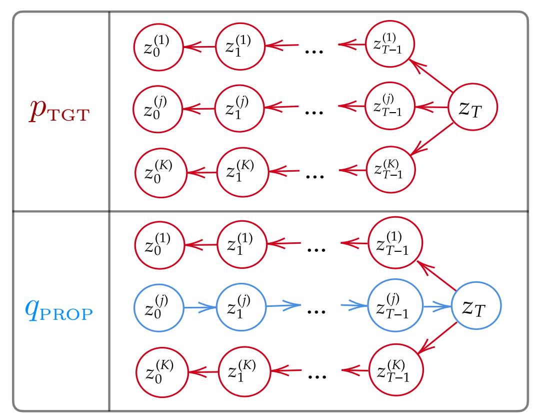
BDMC Lower Bound

(Multi-Sample AIS)

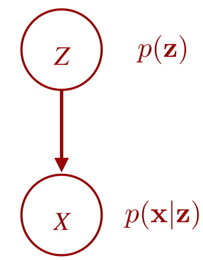


BDMC Upper Bound

(Grosse et al 2015)



Results



- Evaluating MI between latent **z** and generated data **x** in VAEs, GANs

using learned Gaussian $q_\theta(\mathbf{z}|\mathbf{x})$

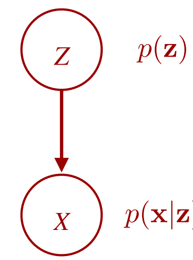
Method	MNIST-VAE10	MNIST-VAE100	MNIST-GAN10	MNIST-GAN100
AIS (T=500)	**(34.16, 34.29)**	**(80.19, 82.34)**	**(21.60, 23.06)**	(**25.58**, 29.53)
AIS (T=30K)	**(34.21, 34.21)**	**(80.77, 80.80)**	**(22.01, 22.01)**	**(26.53, 26.54)**
IWAE (K=1K)	(31.69, **34.24**)	(51.44, 85.30)	(11.14, 52.73)	(10.14, 201.18)
IWAE (K=1M)	**(34.10, 34.22)**	(58.35, 83.39)	(17.76, 30.88)	(16.98, 58.04)

Method	CIFAR-GAN10	CIFAR-GAN100
AIS T=500	(48.16, 136.15)	(145.19, 2786.53)
AIS T=100K	**(72.85, 73.54)**	(479.27, 484.84)
IWAE K=1K	(23.58, **74.00**)	(26.98, 5283.13)
IWAE K=1M	(30.73, **73.36**)	(33.81, 5271.56)

IWAE lower bound improves at rate of *at most* **log K** (# importance samples)

$\Delta \log K = 6.91$

Results



- Evaluating MI between latent **z** and generated data **x** in VAEs, GANs

using learned Gaussian $q_\theta(\mathbf{z}|\mathbf{x})$

Method	MNIST-VAE10	MNIST-VAE100	MNIST-GAN10	MNIST-GAN100
AIS (T=500)	**(34.16, 34.29)**	**(80.19, 82.34)**	**(21.60, 23.06)**	(**25.58**, 29.53)
AIS (T=30K)	**(34.21, 34.21)**	**(80.77, 80.80)**	**(22.01, 22.01)**	**(26.53, 26.54)**
IWAE (K=1K)	(31.69, **34.24**)	(51.44, 85.30)	(11.14, 52.73)	(10.14, 201.18)
IWAE (K=1M)	**(34.10, 34.22)**	(58.35, 83.39)	(17.76, 30.88)	(16.98, 58.04)

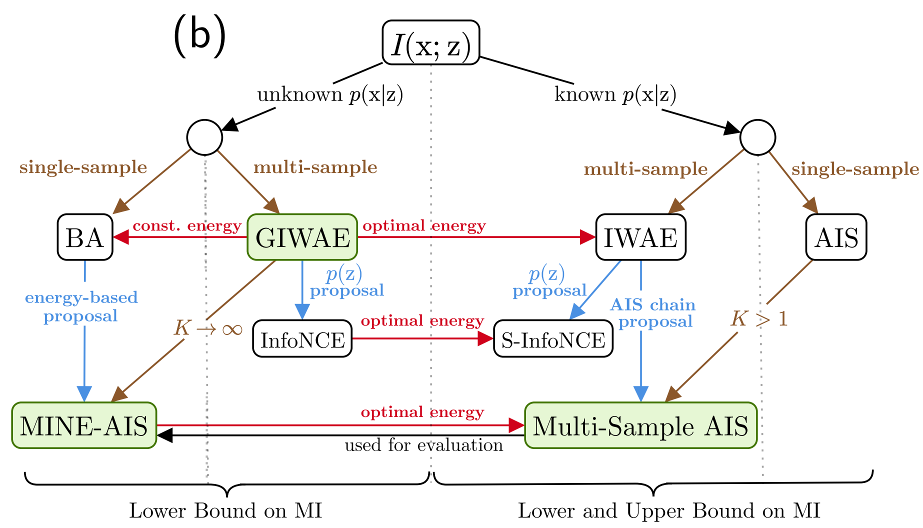
Method	CIFAR-GAN10	CIFAR-GAN100
AIS T=500	(48.16, 136.15)	(145.19, 2786.53)
AIS T=100K	**(72.85, 73.54)**	(479.27, 484.84)
IWAE K=1K	(23.58, **74.00**)	(26.98, 5283.13)
IWAE K=1M	(30.73, **73.36**)	(33.81, 5271.56)

Multi-Sample AIS
sandwich bounds tighten
at rate of **1/T**
(under perfect transitions)

Mutual Information Estimation

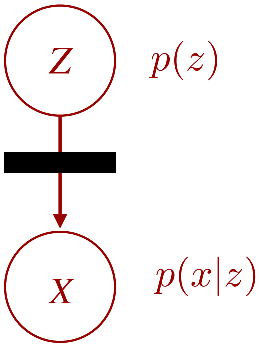
(a) Known $p(\mathbf{x}|\mathbf{z})$, $p(\mathbf{z})$ Multi-Sample Annealed Importance Sampling

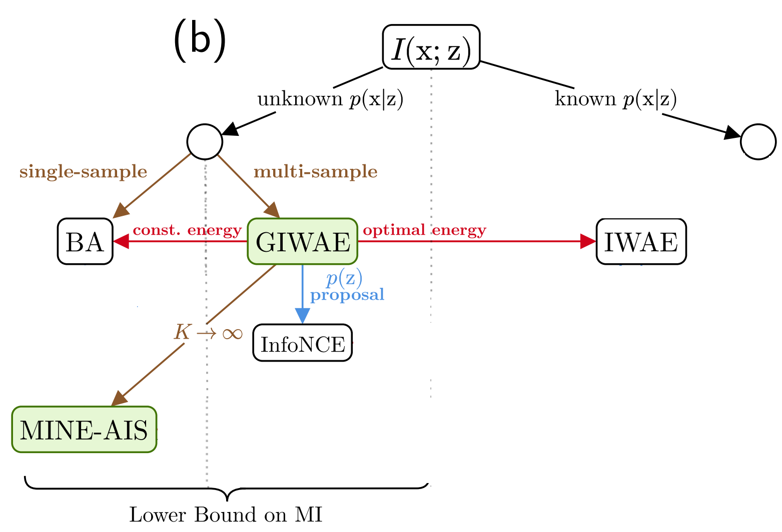
(b) Known $p(\mathbf{z})$ only MINE-AIS, Generalized IWAE



Mutual Information Estimation

(b) Known $p(\mathbf{z})$ only MINE-AIS, Generalized IWAE

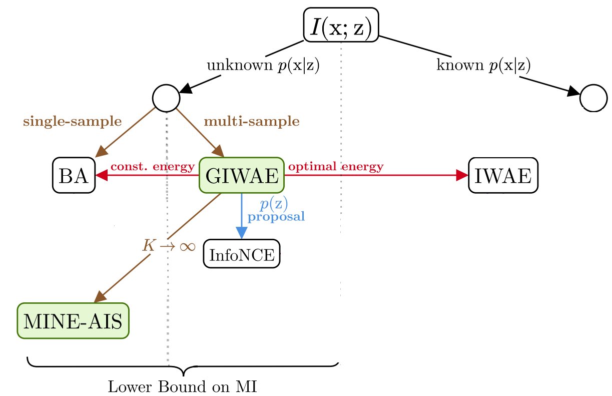




Barber-Agakov Lower Bound

(b) Known $p(\mathbf{z})$ only

$$I_p(X;Z) = \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\big[\log \underbrace{p(\mathbf{z}|\mathbf{x})}_{\text{posterior inference}}\big] - \mathbb{E}_{p(\mathbf{z})}\big[\log p(\mathbf{z})\big]$$



Barber-Agakov Lower Bound

(b) Known $p(\mathbf{z})$ only

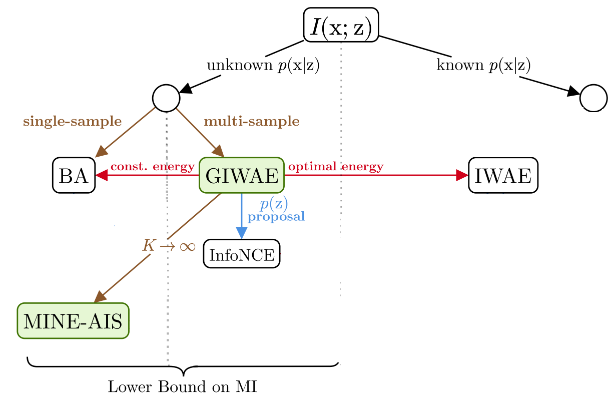
$$I_p(X;Z) = \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log p(\mathbf{z}|\mathbf{x})\right] - \mathbb{E}_{p(\mathbf{z})}\left[\log p(\mathbf{z})\right]$$

$$\geq \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log \frac{q_\theta(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})}\right]$$

Barber-Agakov Lower Bound

(Barber & Agakov 2003)

($\log p(\mathbf{x})$ upper bound with gap $\mathbb{E}_{p(\mathbf{x})}\left[D_{KL}[p(\mathbf{z}|\mathbf{x})||q_\theta(\mathbf{z}|\mathbf{x})]\right]$)



Barber-Agakov Lower Bound

$$I_p(X;Z) \;\geq \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log \frac{q_\theta(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})}\right]$$

Barber-Agakov Lower Bound

(Barber & Agakov 2003)

(gap $\mathbb{E}_{p(\mathbf{x})}\left[D_{KL}[p(\mathbf{z}|\mathbf{x})||q_\theta(\mathbf{z}|\mathbf{x})]\right]$)

InfoNCE

learned critic function

$$I_p(X;Z) \geq \quad \mathbb{E}_{p(\mathbf{x})p(\mathbf{z}^{(1)}|\mathbf{x})\prod_{k=2}^{K}p(\mathbf{z}^{(k)})}\left[\log \frac{e^{T_\phi(\mathbf{x},\mathbf{z}^{(1)})}}{\frac{1}{K}\sum_{k=1}^{K}e^{T_\phi(\mathbf{x},\mathbf{z}^{(k)})}}\right]$$

InfoNCE contrastive lower bound

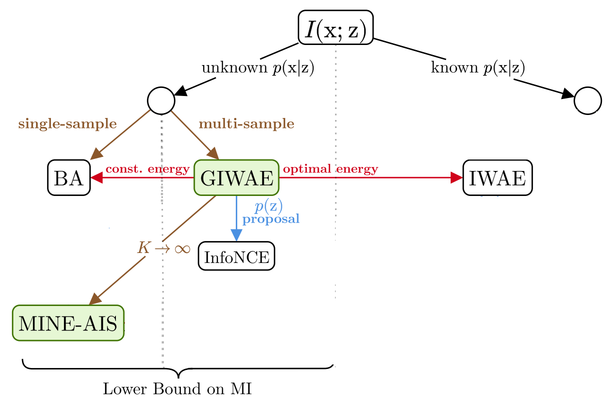
(Poole et. al 2019, van Oord et. al 2018)

Generalized IWAE

negative sampling under q_θ

$$I_p(X;Z) \geq \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log \frac{q_\theta(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})}\right] + \mathbb{E}_{p(\mathbf{x})p(\mathbf{z}^{(1)}|\mathbf{x}) \prod_{k=2}^{K} q_\theta(\mathbf{z}^{(k)}|\mathbf{x})}\left[\log \frac{e^{T_\phi(\mathbf{x},\mathbf{z}^{(1)})}}{\frac{1}{K}\sum_{k=1}^{K} e^{T_\phi(\mathbf{x},\mathbf{z}^{(k)})}}\right]$$

- Special Cases
  - IWAE for $T^*(\mathbf{x},\mathbf{z}) = \log \dfrac{p(\mathbf{x},\mathbf{z})}{q_\theta(\mathbf{z}|\mathbf{x})}$
  - Barber Agakov
  - InfoNCE



GIWAE Infinite Sample Limit

Generalized IWAE lower bound

$$I_p(X;Z) \geq \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log \frac{q_\theta(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})}\right] + \mathbb{E}_{p(\mathbf{x})p(\mathbf{z}^{(1)}|\mathbf{x}) \prod_{k=2}^{K} q_\theta(\mathbf{z}^{(k)}|\mathbf{x})}\left[\log \frac{e^{T_\phi(\mathbf{x},\mathbf{z}^{(1)})}}{\frac{1}{K}\sum_{k=1}^{K} e^{T_\phi(\mathbf{x},\mathbf{z}^{(k)})}}\right]$$

$$\boxed{K \to \infty}$$

MINE-AIS lower bound

$$I_p(X;Z) \geq \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log \frac{q_\theta(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})}\right] + \mathbb{E}_{p(\mathbf{x})p(\mathbf{z}|\mathbf{x})}\left[\log \frac{e^{T_\phi(\mathbf{x},\mathbf{z})}}{\mathbb{E}_{q_\theta(\mathbf{z}|\mathbf{x})}\left[e^{T_\phi(\mathbf{x},\mathbf{z})}\right]}\right]$$

MINE-AIS

- Energy-Based Posterior $\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x}) = \frac{1}{\mathcal{Z}(\mathbf{x})} q_\theta(\mathbf{z}|\mathbf{x}) e^{T_\phi(\mathbf{x},\mathbf{z})}$
- Training : contrastive divergence training of energy-based posterior
- Evaluation : Multi-Sample AIS to evaluate $\log \mathcal{Z}(\mathbf{x}) = \log \mathbb{E}_{q_\theta(\mathbf{z}|\mathbf{x})}\left[e^{T_\phi(\mathbf{x},\mathbf{z})}\right]$

$$I_p(X;Z) \geq \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log \frac{q_\theta(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})}\right] + \mathbb{E}_{p(\mathbf{x})p(\mathbf{z}|\mathbf{x})}\left[\log \frac{e^{T_\phi(\mathbf{x},\mathbf{z})}}{\mathbb{E}_{q_\theta(\mathbf{z}|\mathbf{x})}\left[e^{T_\phi(\mathbf{x},\mathbf{z})}\right]}\right]$$

Lower Bound Results

optimal T_ϕ

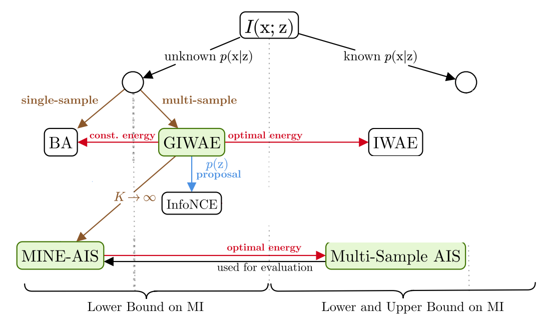
$K \to \infty$

optimal T_ϕ

energy-based posterior

	Information Used	MNIST VAE-20	MNIST GAN-20
USING (MULTI-SAMPLE AIS) $I(\mathbf{x};\mathbf{z})$	$p(\mathbf{x},\mathbf{z})$	65.1	53.4
MINE-AIS (q_θ, T_ϕ)	$p(\mathbf{z})$	57.7	40.8
IWAE (q_θ, K)	$p(\mathbf{x},\mathbf{z})$ $K = 1000$	45.1	27.9
GIWAE (q_θ, T_ϕ, K)	$p(\mathbf{z})$ $K = 1000$	43.7	27.2
BARBER-AGAKOV (q_θ)	$p(\mathbf{z})$	37.9	21.4

using learned Gaussian $q_\theta(\mathbf{z}|\mathbf{x})$



Thanks!