

성능은 좋지만 sampling이 느린 EBM과 빠르지만 설명하지 않은 VAE를 합침.

VAE는 data distribution의 전기 structure를 험해하고, EBM으로 수렴.

▶ Introduction

EBM과 VAE는 모두 architecture of 2인자이 같다.

그러면 VAE의 경우 낮은 data density의 높은 확률로 활동하는 경향이 있고, 이는 이상한 data 풍경의 실패의 원인이 된다.

EBM의 경우 out-of-distribution을 잘 감지하지만, MCMC 고정점의 시간, 계산 소모가 있다.

그러나 두 model의 결점만 개선한 VAEBM을 제작.

Contribution

1. VAE의 data space에서 EBM을 제작.

2. VAE의 EBM 훈련의 decompose

3. VAEBM은 MCMC sampling이 VAE의 latent space로 비워지는 것

4. SOTA.

* Background.

- Energy-based Model

$E_\psi(x)$ 이 ψ 의 대신 energy function²을 의미하고, Z_ψ 는 partition function (normalize) 일 때,

EBM은 $P_\psi(x) = \exp(-E_\psi(x)) / Z_\psi$ 의 Gibbs distribution을 가정한다.

data distribution $P_{data}(x)$ 의 대신 NLL은 $L(\psi) = -E_{x \sim P_d(x)} [\log P_\psi(x)]$

↳ 미분의 성질, $\partial_\psi L(\psi) = \mathbb{E}_{x \sim p_d(x)} [-\partial_\psi E_\psi(x)] + \mathbb{E}_{x \sim p_\psi(x)} [\partial_\psi E_\psi(x)]$

↳ intractable or MCMC.

→ 보통 Langevin Dynamics 가 사용됨

$$x_{t+1} = x_t - \frac{\eta}{2} \nabla_x E_\psi(x_t) + \sqrt{\eta} \omega_t, \quad \omega_t \sim \mathcal{N}(0, I),$$

- EBM의 추적적인 설명 (Notes On Contrastive Divergence)

f 가 pdf 또는 cdf고, p 는 probability 일 때,

$$P(x; \theta) = \frac{1}{Z(\theta)} f(x; \theta), \quad Z(\theta) = \int f(x; \theta) dx$$

훈련 set에 대한 학률은 $P(X; \theta) = \prod_{k=1}^K \frac{1}{Z(\theta)} f(x_k; \theta)$

미니배치 표본으로는 E 는 $E(X; \theta) = \log Z(\theta) - \frac{1}{K} \sum_{k=1}^K \log f(x_k; \theta) \quad (= -\log P(x; \theta))$

예를들어, $N(0, I)$ 을 따르고, x 가 1-dim일 땐, $Z(\theta)=1$ 로 constant 계산이 가능하다.

p 가 여러 N 의 합일 땐, $\log Z(\theta)$ 는 $\log N$ 으로 계산할 수 있지만, 서로 dependent 때에는 optimal param을 한 번에 찾을 수 없다.

따라서 gradient descent 를 사용해야 함

p 가 여러 dim의 N 일 땐, $f(x; \theta) = \prod_{i=1}^n N(x_i; \mu_i, \sigma_i)$

$Z(\theta)$ 는 constant가 아니고, 계산하기가 불가능하다.

Contrastive Divergence (이하 CD)로 이를 해결 가능

E의 기울기를 추정하면,

$$\frac{\partial E(x; \theta)}{\partial \theta} = \frac{\partial \log Z(\theta)}{\partial \theta} - \frac{1}{K} \sum_{i=1}^K \frac{\partial \log f(x_i; \theta)}{\partial \theta}$$
$$\frac{\partial \log Z(\theta)}{\partial \theta} - \left\langle \frac{\partial \log f(x; \theta)}{\partial \theta} \right\rangle_x$$

$\frac{\partial \log Z(\theta)}{\partial \theta}$ 의 정의.

$$\frac{\partial \log Z(\theta)}{\partial \theta} = \frac{1}{Z(\theta)} \frac{\partial Z(\theta)}{\partial \theta}$$
$$= \frac{1}{Z(\theta)} \frac{\partial}{\partial \theta} \int f(x; \theta) dx$$
$$= \frac{1}{Z(\theta)} \int \frac{\partial f(x; \theta)}{\partial \theta} dx$$
$$= \frac{1}{Z(\theta)} \int f(x; \theta) \frac{\partial \log f(x; \theta)}{\partial \theta} dx$$
$$= \int p(x; \theta) \frac{\partial \log f(x; \theta)}{\partial \theta} dx$$
$$= \left\langle \frac{\partial \log f(x; \theta)}{\partial \theta} \right\rangle_{p(x; \theta)}$$

이로 인해, 초기된 $p(x; \theta)$ 를 통해 numerically approximation이 가능하다.

근을 찾기 때문에 바로 $p(x; \theta)$ 를 넣을 수 있기 때문에, MCMC로 proposed distribute를 target distribute로 만든다.

다면 MCMC 일 때, $x^n \therefore x^0 = x$

$$\frac{\partial E(x; \theta)}{\partial \theta} = \left\langle \frac{\partial \log f(x; \theta)}{\partial \theta} \right\rangle_{x^\infty} - \left\langle \frac{\partial \log f(x; \theta)}{\partial \theta} \right\rangle_{x^0}$$

정확한 target의 경우 $\theta \rightarrow \infty$ 일 때, derivate의 경우, 몇 가지로도 근사가 가능하다.

설마 $\theta = 1$ 일 때도, 경험적으로 ML 학습이 가능

$$\text{따라서, } \theta_{t+1} = \theta_t + \eta \left(\left\langle \frac{\partial \log f(x; \theta)}{\partial \theta} \right\rangle_{x^0} - \left\langle \frac{\partial \log f(x; \theta)}{\partial \theta} \right\rangle_{x^t} \right)$$

* Energy-based Variational AutoEncoder.

VAE의 가장 큰 문제는 높은 (low density spaces) 큰 probability를 할당하는 것이다.

이를 해결하기 위해 VAE로 생성하고, data space에서 EBM을 적용한 VAEBM을 제시한다.

$$\partial_\psi L(\psi) = \mathbb{E}_{x \sim p_d(x)} [-\partial_\psi E_\psi(x)] + \mathbb{E}_{x \sim p_\psi(x)} [\partial_\psi E_\psi(x)]$$

미지 우량의 첫번째는 positive, 두번째는 negative phase라고 한다.

negative phase는 model가 normal sampling 시에 data distribution에 맞지 않은 sample을 생성할 수 있다.

이는 E function이 으뜸 명시적으로 간소된다.

pixel space에서 수행된 E function은, GAN의 Det 를 봄고, 이를 성능 좋은 sample을 생성한다.

이제正规화된 VAEBM은 generative model은

$$h_{\psi,\theta}(x, z) = \frac{1}{Z_{\psi,\theta}} P_\theta(x, z) e^{-E_\psi(x)}$$

\hookrightarrow VAE를 의미

$$\hookrightarrow P_\theta(x, z) = P_\theta(z) P_\theta(x|z), E_\psi(x) \text{는 } E \text{ function}, Z_{\psi,\theta} = \int P_\theta(x) e^{-E_\psi(x)} dx$$

$$h_{\psi,\theta}(x) = \frac{1}{Z_{\psi,\theta}} \int p_\theta(x, z) e^{-E_\psi(x)} dz = \frac{1}{Z_{\psi,\theta}} p_\theta(x) e^{-E_\psi(x)}.$$

\hookrightarrow marginalizing

VAEBM은 marginal log-likelihood를 최대화하도록 함

$$\log h_{\psi,\theta}(x) = \log p_\theta(x) - E_\psi(x) - \log Z_{\psi,\theta} \quad (5)$$

$$\geq \underbrace{\mathbb{E}_{z \sim q_\phi(z|x)} [\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p(z))}_{\mathcal{L}_{\text{vae}}(x, \theta, \phi)} - E_\psi(x) - \underbrace{\log Z_{\psi,\theta}}_{\mathcal{L}_{\text{EBM}}(x, \psi, \theta)}, \quad (6)$$

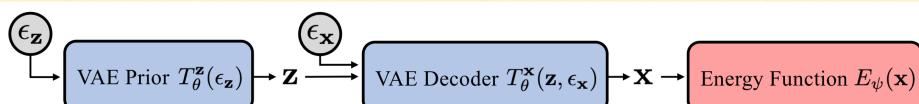


Figure 1: Our VAEBM is composed of a VAE generator (including the prior and decoder) and an energy function that operates on samples x generated by the VAE. The VAE component is trained first, using the standard VAE objective; then, the energy function is trained while the generator is fixed. Using the VAE generator, we can express the data variable x as a deterministic function of white noise samples e_z and e_x . This allows us to reparameterize sampling from our VAEBM by sampling in the joint space of e_z and e_x . We use this in the negative training phase (see Sec. 3.1).

- Training

식 600M Z의 gradient는

→ 흥영은 Appendix A

$$\partial_\psi \log Z_{\psi, \theta} = \mathbb{E}_{\mathbf{x} \sim h_{\psi, \theta}(\mathbf{x}, \mathbf{z})} [-\partial_\psi E_\psi(\mathbf{x})] \rightarrow \text{MCMC 이용}$$

$$\partial_\theta \log Z_{\psi, \theta} = \mathbb{E}_{\mathbf{x} \sim h_{\psi, \theta}(\mathbf{x}, \mathbf{z})} [\partial_\theta \log p_\theta(\mathbf{x})]$$

↳ $\frac{\partial}{\partial \theta} \log p_\theta(\mathbf{x})$ 는 Intractable.

↳ VAE의 posterior sampling

↳ Appendix A

계산 복잡성을 고려해 두 단계로 VAEBM 설계.

stage 1 단계는 $L_{VAE}(\mathbf{x}, \theta, \phi)$ 만 훈련,

stage 2 단계 VAE model은 fixed로, EBM 만 훈련.

↳ θ 가 fix 되면 때문에, $L_{EBM}(\mathbf{x}, \psi, \theta)$ 단계가 만 고려되면서 단계.

$$\rightarrow \partial_\psi L(\psi) = \mathbb{E}_{\mathbf{x} \sim p_d(\mathbf{x})} [-\partial_\psi E_\psi(\mathbf{x})] + \mathbb{E}_{\mathbf{x} \sim h_{\psi, \theta}(\mathbf{x}, \mathbf{z})} [\partial_\psi E_\psi(\mathbf{x})]$$

$$\hookrightarrow L(\psi) = \mathbb{E}_{\mathbf{x} \sim p_d(\mathbf{x})} [L_{EBM}(\mathbf{x}, \psi, \theta)]$$

↳ positive, negative phase로 분할됨.

* Reparameterized sampling in the negative phase

보통 negative phase 단계 MCMC를 사용하고, 여러 인수가 있으면 반복적으로 사용할 수 있지만, 논문에서는 다른 방법을 사용

\mathbf{x}, \mathbf{z} 에 대한 reparameterized trick을 사용하여, sampling을 가능.

VAE 단계 $\mathbf{z} \sim p(\mathbf{z})$ 후 $\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{z})$ 를 한다.

여기에서, 그는 $\mu + \epsilon (\sim N(0, I))$ 로 reparameterize가 가능했다.

↳ noise ϵ 를 sample로 바꾸는 변환

이후 같이, $(\epsilon_z, \epsilon_x) \sim p_\epsilon = N(0, I)$ 을 고정하고, deterministic transformation T_θ 로 sampling 할 수 있다.

$$\mathbf{z} = T_\theta^z(\epsilon_z), \quad \mathbf{x} = T_\theta^x(\mathbf{z}(\epsilon_z), \epsilon_x) = T_\theta^x(T_\theta^z(\epsilon_z), \epsilon_x).$$

$$h_{\psi, \theta}(\epsilon_x, \epsilon_z) \propto e^{-E_\psi(T_\theta^x(T_\theta^z(\epsilon_z), \epsilon_x))} p_\epsilon(\epsilon_x, \epsilon_z), \rightarrow \text{Appendix B}$$

↳ 반복적인 MCMC보다 빠르고, Langevin Dynamics처럼 hyperparam (step size)를 설정할 필요가 없다.

* The advantage of two-stage training

$\log Z_{\theta, \phi}$ 의 복잡한 계산을 피하는 것 이외에도 추가적인 이점이 있다.

stage 1에서 data와 gen의 거리를 향하고,

stage 2에서 mismatch를 줄인다.

pre-trained VAE $P_\theta(x)$ 이미 이미 $P_d(x)$ 와 같은 근사를 했기 때문에

작은 update만 필요하다.

그리고 data distribution보다 low-dim이고 smooth한 latent space로, 효과적인 MCMC가 가능하다.

stage 2에서 VAE를 훈련하는 extension은 Appendix C.

* Related Work

* Experiments.

VAE는 NVAE를 뜻, EBM은 simple Residual

priorer data는 Gaussian, ICMH LDZ sampling

- Image Generation.

Table 1: IS and FID scores for unconditional generation on CIFAR-10.

	Model	IS↑	FID↓
Ours	VAEBM w/o persistent chain	8.21	12.26
	VAEBM w/ persistent chain	8.43	12.19
EBMs	IGEBM (Du & Mordatch, 2019)	6.02	40.58
	EBM with short-run MCMC (Nijkamp et al., 2019b)	6.21	-
	F-div EBM (Yu et al., 2020a)	8.61	30.86
	FlowCE (Gao et al., 2020)	-	37.3
	FlowEBM (Nijkamp et al., 2020)	-	78.12
	GEBM (Arbel et al., 2020)	-	23.02
Other Likelihood Models	Divergence Triangle (Han et al., 2020)	-	30.1
	Glow (Kingma & Dhariwal, 2018)	3.92	48.9
	PixelCNN (Oord et al., 2016b)	4.60	65.93
	NVAE (Vahdat & Kautz, 2020)	5.51	51.67
Score-based Models	VAE with EBM prior (Pang et al., 2020)	-	70.15
	NCSN (Song & Ermon, 2019)	8.87	25.32
	NCSN v2 (Song & Ermon, 2020)	-	31.75
	Multi-scale DSM (Li et al., 2019)	8.31	31.7
GAN-based Models	Denoising Diffusion (Ho et al., 2020)	9.46	3.17
	SNGAN (Miyato et al., 2018)	8.22	21.7
	SNGAN+DDLS (Che et al., 2020)	9.09	15.42
	SNGAN+DCD (Song et al., 2020)	9.11	16.24
	BigGAN (Brock et al., 2018)	9.22	14.73
Others	StyleGAN2 w/o ADA (Karras et al., 2020a)	8.99	9.9
	PixelIQN (Ostrovski et al., 2018)	5.29	49.46
	MoLM (Ravuri et al., 2018)	7.90	18.9

Table 2: Generative performance on CelebA 64

Model	FID↓
VAEBM (ours)	5.31
NVAE (Vahdat & Kautz)	14.74
Flow CE (Gao et al.)	12.21
Divergence Triangle (Han et al.)	24.7
NCSNv2 (Song & Ermon)	26.86
COCO-GAN (Lin et al.)	4.0
QA-GAN (Parimala & Channappayya)	6.42

Table 3: Generative performance on CelebA HQ 256

Model	FID↓
VAEBM (ours)	20.38
NVAE (Vahdat & Kautz)	45.11
GLOW (Kingma & Dhariwal)	68.93
Advers. LAE (Pidhorskyi et al.)	19.21
PGGAN (Karras et al.)	8.03

-Ablation study

Table 4: Comparison for IS and FID on CIFAR-10 between several related training methods.

Model	IS↑	FID↓
NVAE (Vahdat & Kautz)	5.19	55.97
EBM on \mathbf{x} (Du & Mordatch)	5.85	48.89
EBM on \mathbf{x} , MCMC init w/ NVAE	7.28	29.32
WGAN w/ NVAE decoder	7.41	20.39
VAEBM (ours)	8.15	12.96

Table 5: Mode coverage on StackedMNIST.

Model	Modes↑	KL↓
VEEGAN (Srivastava et al.)	761.8	2.173
PacGAN (Lin et al.)	992.0	0.277
PresGAN (Dieng et al.)	999.6	0.115
InclusiveGAN (Yu et al.)	997	0.200
StyleGAN2 (Karras et al.)	940	0.424
VAEBM (ours)	1000	0.087

- Test for Spurious or Missing Model.

Train에 Test로 따른 kl-divergence.

↳ Overfit하지 않음.

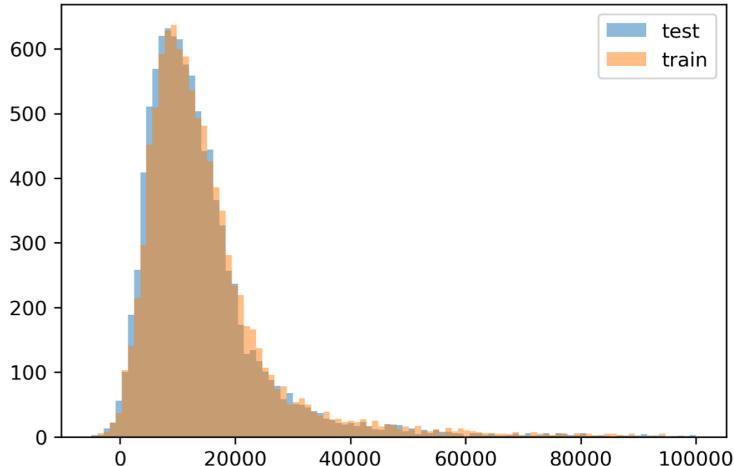


Figure 6: Histogram of unnormalized log-likelihoods on 10k CIFAR-10 train and test set images

OOD 모니터링 probability를 측정(AUROC)과 같은지 측정

Table 6: Table for AUROC↑ of $\log p(\mathbf{x})$ computed on several OOD datasets. In-distribution dataset is CIFAR-10. Interp. corresponds to linear interpolation between CIFAR-10 images.

		SVHN	Interp.	CIFAR100	CelebA
Unsupervised Training	NVAE (Vahdat & Kautz, 2020)	0.42	0.64	0.56	0.68
	Glow (Kingma & Dhariwal, 2018)	0.05	0.51	0.55	0.57
	IGEBM (Du & Mordatch, 2019)	0.63	0.7	0.5	0.7
	Divergence Triangle (Han et al., 2020)	0.68	-	-	0.56
	VAEBM (ours)	0.83	0.7	0.62	0.77
Supervised Training	JEM (Grathwohl et al., 2020a)	0.67	0.65	0.67	0.75
	HDGE (Liu & Abbeel, 2020)	0.96	0.82	0.91	0.8

- Exact likelihood estimate on 2D Toy Data.

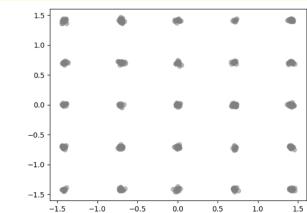
VAEBM은 명백한 likelihood 모델이다.

다른 likelihood 모델과 같이, partition function $\log Z(\theta)$ 때문에 정확한 likelihood 추정이 힘들다.

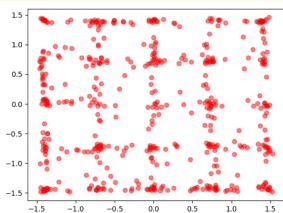
AIS(Ampled Importance Sampling)이 가능하지 않거나 고차원의 경우, high-dimensional 시간이 더욱 오래 걸림.

또한 AIS는 $\log Z$ 의 lower bound이 \rightarrow 비교를 어렵게 한다.

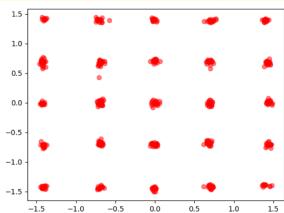
VAEBM의 trained distribution을 개선하는 능력을 통해 2D Toy data의 샘플링.



(a) Samples from the true distribution



(b) Samples from VAE



(c) Samples from VAEBM

Figure 4: Qualitative results on the 25-Gaussians dataset

- Sampling Efficient.

1000번 정도 sampling하는데 NCSN보다 10배 빠름

* Conclusion.

VAE+EBM이 공정한 VAEBM을 제시.

VAE의 embedding space에 sampling 가능.

↳ 효율적 훈련, sampling.

* Appendix.

A. Deriving the gradient of $\log Z_{\psi, \theta}$

Recall that $Z_{\psi, \theta} = \int p_{\theta}(\mathbf{x}) e^{-E_{\psi}(\mathbf{x})} d\mathbf{x}$. For the derivative of $\log Z_{\psi, \theta}$ w.r.t. θ , we have:

$$\begin{aligned}\frac{\partial}{\partial \theta} \log Z_{\psi, \theta} &= \frac{\partial}{\partial \theta} \log \left(\int p_{\theta}(\mathbf{x}) e^{-E_{\psi}(\mathbf{x})} d\mathbf{x} \right) = \frac{1}{Z_{\psi, \theta}} \int \frac{\partial p_{\theta}(\mathbf{x})}{\partial \theta} e^{-E_{\psi}(\mathbf{x})} d\mathbf{x} \\ &= \frac{1}{Z_{\psi, \theta}} \int p_{\theta}(\mathbf{x}) \frac{\partial \log p_{\theta}(\mathbf{x})}{\partial \theta} d\mathbf{x} = \int h_{\psi, \theta}(\mathbf{x}) \frac{\partial \log p_{\theta}(\mathbf{x})}{\partial \theta} d\mathbf{x} \\ &= \mathbb{E}_{\mathbf{x} \sim h_{\psi, \theta}(\mathbf{x}, \mathbf{z})} \left[\frac{\partial \log p_{\theta}(\mathbf{x})}{\partial \theta} \right]\end{aligned}\quad (10)$$

마찬가지로,

$$\frac{\partial}{\partial \psi} \log Z_{\psi, \theta} = \mathbb{E}_{\mathbf{x} \sim h_{\psi, \theta}(\mathbf{x}, \mathbf{z})} \left[-\frac{\partial E_{\psi}(\mathbf{x})}{\partial \psi} \right]$$

Eq 10을 흡장하면,

$$\frac{\partial}{\partial \theta} \log Z_{\psi, \theta} = \mathbb{E}_{\mathbf{x} \sim h_{\psi, \theta}(\mathbf{x}, \mathbf{z})} \left[\mathbb{E}_{\mathbf{z}' \sim p_{\theta}(\mathbf{z}' | \mathbf{x})} \left[\frac{\partial \log p_{\theta}(\mathbf{x}, \mathbf{z}')}{\partial \theta} \right] \right]$$

↳ MCMC

x의 posterior, $p_{\theta}(\mathbf{z}' | \mathbf{x})$ 로 대체 가능

↳ ϕ 의 성능에 달려있음

↳ 복잡성↑

↳ 따라서 MCMC로 $p_{\theta}(\mathbf{z}' | \mathbf{x})$ sampling

↳ 속도를 높이기 위해 일부 $(\mathbf{x}, \mathbf{z}) \sim h_{\psi, \theta}(\mathbf{x}, \mathbf{z})$ 로 부터 z 추적.

↳ 이를면 $(\mathbf{x}, \mathbf{z}) \sim h_{\psi, \theta}(\mathbf{x}, \mathbf{z})$ 와 $\mathbf{z}' \sim p_{\theta}(\mathbf{z}' | \mathbf{x})$ 로 같은 두번 MCMC

↳ ∴ VAE를 고성능 2 stage 사용

B. Reparameterization for EBM

$(\epsilon_x, \epsilon_z) \sim p_\epsilon(\epsilon_x, \epsilon_z)$ 일 때,

$$T_\theta(\epsilon_x, \epsilon_z) = (T_\theta^x(T_\theta^z(\epsilon_z), \epsilon_x), T_\theta^z(\epsilon_z)) = (\mathbf{x}, \mathbf{z}).$$

$$p_\theta(\mathbf{x}, \mathbf{z}) = p_\epsilon(T_\theta^{-1}(\mathbf{x}, \mathbf{z})) \left| \det \left(J_{T_\theta^{-1}}(\mathbf{x}, \mathbf{z}) \right) \right|,$$

$\mathbf{z} \sim \mathcal{N}(\mu_z, \sigma_z)$ 일 때, $\mathbf{x} | \mathbf{z} \sim \mathcal{N}(\mu_x(\mathbf{z}), \sigma_x(\mathbf{z}))$ 일 때.

$$\mathbf{z} = T_\theta^z(\epsilon_z) = \mu_z + \sigma_z \cdot \epsilon_z, \quad \mathbf{x} = T_\theta^x(\epsilon_x, \epsilon_z) = \mu_x(\mathbf{z}) + \sigma_x(\mathbf{z}) \cdot \epsilon_x,$$

$$J_{T_\theta^{-1}}(\mathbf{x}, \mathbf{z}) = [\sigma_x(\mathbf{z})^{-1}, \sigma_z^{-1}].$$

따라서, EBM 모델,

$$h_{\psi, \theta}(\mathbf{x}, \mathbf{z}) = \frac{e^{-E_\psi(\mathbf{x})} p_\theta(\mathbf{x}, \mathbf{z})}{Z_{\psi, \theta}}.$$

$$h_{\psi, \theta}(\epsilon_x, \epsilon_z) = h_{\psi, \theta}(T_\theta(\epsilon_x, \epsilon_z)) |\det(J_{T_\theta}(\epsilon_x, \epsilon_z))|, \quad \text{즉각적이 표현 가능}$$

다음과 같이 증명 가능.

$$h_{\psi, \theta}(\epsilon_x, \epsilon_z) \propto e^{-E_\psi(T_\theta^x(T_\theta^z(\epsilon_z), \epsilon_x))} p_\epsilon(\epsilon_x, \epsilon_z),$$

다음과 같이 증명 가능.

$$h_{\psi, \theta}(\epsilon_x, \epsilon_z) = h_{\psi, \theta}(T_\theta(\epsilon_x, \epsilon_z)) |\det(J_{T_\theta}(\epsilon_z, \epsilon_x))| \quad (16)$$

$$= \frac{1}{Z_{\psi, \theta}} e^{-E_\psi(T_\theta(\epsilon_x, \epsilon_z))} p_\theta(T_\theta(\epsilon_x, \epsilon_z)) |\det(J_{T_\theta}(\epsilon_x, \epsilon_z))| \quad (17)$$

$$= \frac{1}{Z_{\psi, \theta}} e^{-E_\psi(T_\theta(\epsilon_x, \epsilon_z))} p_\epsilon(T_\theta^{-1}(\mathbf{x}, \mathbf{z})) \left| \det \left(J_{T_\theta^{-1}}(\mathbf{x}, \mathbf{z}) \right) \right| |\det(J_{T_\theta}(\epsilon_x, \epsilon_z))| \quad (18)$$

$$= \frac{1}{Z_{\psi, \theta}} e^{-E_\psi(T_\theta(\epsilon_x, \epsilon_z))} p_\epsilon(T_\theta^{-1}(\mathbf{x}, \mathbf{z})) \quad (19)$$

$$= \frac{1}{Z_{\psi, \theta}} e^{-E_\psi(T_\theta(\epsilon_x, \epsilon_z))} p_\epsilon(\epsilon_x, \epsilon_z), \quad (20)$$

B.1. Comparison of Sampling in (E_x, E_z) -space and in (x, z) -space.

