Energy-Based Probability Estimation with Variational Ancestral Langevin Sampler

Jianwen Xie, Zilong Zheng, Ping Li

October 15, 2020

Energy-Based Generative Models

Energy-based Model

Let x be an input image, $U_\theta(x)$ be an energy function where θ is a set of trainable parameters, an EBM is defined as an unnormalized probability density:

$$p_\theta(x) = \frac{1}{Z(\theta)} \exp[-U_\theta(x)], \tag{1}$$

where $Z(\theta) = \int \exp[-U_\theta(x)]dx$ is a normalizing constant.

We study the energy-based generative model whose energy function $U_\theta(x)$ is parameterized by a non-linear function, e.g., ConvNet. [1]

[1] Jianwen Xie*, Yang Lu*, Song-Chun Zhu, Ying Nian Wu. A Theory of Generative ConvNet. ICML 2016.

MLE by Analysis by Synthesis

MLE Training

Suppose we have a training dataset $D = \{x_i, i = 1, ..., n\}$ and we assume each datapoint is sampled from an unknown distribution p_{data}. The maximum likelihood is to minimize the NLL of the observed data by gradient-based optimization

$$\frac{\partial}{\partial \theta}\mathrm{KL}(p_{\mathrm{data}}(x) || p_\theta(x)) = \mathrm{E}_{x \sim p_{\mathrm{data}}(x)}\left[\frac{\partial U_\theta(x)}{\partial \theta}\right] - \mathrm{E}_{\tilde{x} \sim p_\theta(x)}\left[\frac{\partial U_\theta(\tilde{x})}{\partial \theta}\right] \quad (2)$$

where $\mathrm{E}_{\tilde{x} \sim p_\theta(x)}\left[\frac{\partial U_\theta(\tilde{x})}{\partial \theta}\right]$ is analytically intractable and has to be approximated by MCMC sampling (e.g. Langevin Dynamics).

A gradient-based MCMC: Langevin dynamics

Langvein Dynamics Sampler

Given current energy function $U_\theta(x)$, the initial state $\tilde{x}_0 \sim N(0, I_D)$, the Langevin Dynamics iteratively revises $\tilde{x}$ by finite Langevin steps. For time step t, step size δ, $\tilde{x}_t$ is updated by

$$\tilde{x}_{t+1} = \tilde{x}_t - \frac{\delta^2}{2} \frac{\partial U_\theta(\tilde{x}_t)}{\partial \tilde{x}_t} + \delta N(0, I_D) \tag{3}$$

Challenge

- MCMC is **computationally expensive** and hard to converge.
- Target distribution may have multiple modes separated by low probability regions.
- Long-run MCMC chains are easily **get trapped by local modes**.

EBM with Ancestral Langevin Sampler

Ancestral Langevin Sampling [2]

For the efficient MCMC convergence, we bring in a directed latent variable model $g_\alpha(z)$ to serve as a fast non-iterative sampler to initialize the MCMC sampler.

$$\text{(i)} z \sim N(0, I_d), \hat{x} = g_\alpha(z) + \epsilon \tag{4}$$

where $\hat{x}$ is the initial example generated by ancestral sampling. The goal of $g_\alpha(z)$ is to pursue a good starting point for MCMC sampling, i.e. mimic the the distribution of $p_\theta(x)$.

$$\text{(ii) } \tilde{x}_{t+1} = \tilde{x}_t - \frac{\delta^2}{2}\frac{\partial U_\theta(\tilde{x}_t)}{\partial \tilde{x}} + \delta \mathrm{N}(0, I_D),\ \tilde{x}_0 = \hat{x}, \tag{5}$$

[2] Jianwen Xie, Yang Lu, Ruiqi Gao, Song-Chun Zhu, Ying Nian Wu. Cooperative Training of Descriptor and Generator Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 2018

EBM with Ancestral Langevin Sampler

With $\{\tilde{x}_i\}_{i=1}^{\tilde{n}} \sim p_\theta(x)$, we can compute the gradient in Eq. (2) by

$$\frac{\partial}{\partial\theta}\mathrm{KL}(p_{\mathrm{data}}(x)||p_\theta(x)) \approx \frac{1}{n}\sum_{i=1}^{n}\frac{\partial U_\theta(x_i)}{\partial\theta} - \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\frac{\partial U_\theta(\tilde{x}_i)}{\partial\theta} \tag{6}$$

to update the parameters of EBM.

Now the question is how we learn the latent variable model $q_\alpha(x)$? What strategy?

How to train the latent variable model?

Maximum likelihood estimation of the latent variable model

Given the latent variable model as below

$$z \sim N(0, I_d), \hat{x} = g_\alpha(z) + \epsilon \tag{7}$$

The marginal distribution of $x \sim q_\alpha(x)$ is defined by

$$q_\alpha(x) = \int q_\alpha(x|z)q(z)dz \tag{8}$$

where prior distribution $q(z) = N(0, I_d)$ and conditional distribution $q_\alpha(x|z) = N(g_\alpha(z), \sigma^2 I_D)$. Both posterior distribution $q_\alpha(z|x)$ and marginal distribution $q_\alpha(x)$ are analytically intractable.

How to train the latent variable model?

Alternative Back-propagation (ABP)

ABP maximizes the log-likelihood, whose gradient is

$$\frac{\partial}{\partial\alpha}\mathrm{KL}(p_{\mathrm{data}}(x)||q_\alpha(x)) = \mathrm{E}_{p_{\mathrm{data}}(x)q_\alpha(z|x)}\left[-\frac{\partial}{\partial\theta}\log q_\alpha(z,x)\right]. \quad (9)$$

Variational Auto-Encoder (VAE)

VAE approximates $q_\alpha(z|x)$ by a tractable inference network, e.g., $\pi_\beta(z|x) \sim \mathrm{N}(\mu_\beta(x), \mathrm{diag}(v_\beta(x)))$. The objective of VAE tries to find α and β to minimize

$$\begin{aligned}&\mathrm{KL}(p_{\mathrm{data}}(x)\pi_\beta(z|x)||q_\alpha(z,x))\\ =&\mathrm{KL}(p_{\mathrm{data}}(x)||q_\alpha(x)) + \mathrm{KL}(\pi_\beta(z|x)||q_\alpha(z|x)),\end{aligned} \quad (10)$$

EBM with Ancestral Langevin Sampler

With $\{\tilde{x}_i\}_{i=1}^{\tilde{n}} \sim p_\theta(x)$, we can compute the gradient in Eq. (2) by

$$\frac{\partial}{\partial\theta}\mathrm{KL}(p_{\mathrm{data}}(x)||p_\theta(x)) \approx \frac{1}{n}\sum_{i=1}^{n}\frac{\partial U_\theta(x_i)}{\partial\theta} - \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\frac{\partial U_\theta(\tilde{x}_i)}{\partial\theta} \tag{11}$$

to update the parameters of EBM.

Now the question is how we learn the latent variable model $q_\alpha(x)$?

Q:What strategy?

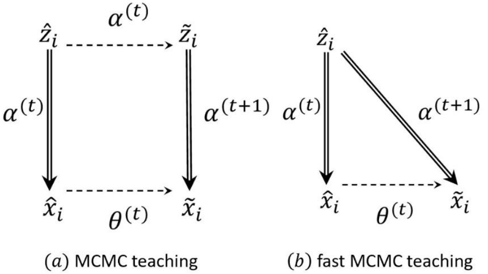
A: MCMC teaching [3]: We train the $q_\alpha(x)$ from the synthesized examples $\{\tilde{x}_i\}$.

---

[3] Jianwen Xie, Yang Lu, Ruiqi Gao, Song-Chun Zhu, Ying Nian Wu. Cooperative Training of Descriptor and Generator Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 2018

Cooperative learning and MCMC teaching

In the original MCMC teaching paper [4], the $q_\alpha(x)$ is trained by ABP from $\{\tilde{x}_i\}$. The resulting model is called Cooperative Networks (CoopNets).



[4] Jianwen Xie, Yang Lu, Ruiqi Gao, Song-Chun Zhu, Ying Nian Wu. Cooperative Training of Descriptor and Generator Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 2018

Variational MCMC teaching

In the proposed framework, we want to learn $q_\alpha(x)$ by VAE from $\{\tilde{x}_i\}$.

To retrieve the latent variable of $\{\tilde{x}_i\}$, we bring in a tractable approximate inference network $\pi_\beta(z|x)$ and infer $z \sim \pi_\beta(z|\tilde{x})$. Then the learning of $\pi_\beta(z|x)$ and $q_\alpha(x|z)$ forms a VAE that treats $\{\tilde{x}_i\}$ as training examples. We call this the variational MCMC teaching [5].

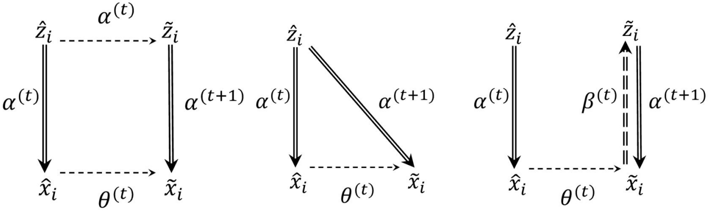
Variational MCMC teaching

Suppose we have $\{\tilde{x}_i\}_{i=1}^{\tilde{n}} \sim \mathcal{M}_{\theta_t} q_{\alpha_t}(x)$ at iteration t, (Let M_θ be the transition kernel of the finite-step MCMC that samples from $p_{\theta_t}(x)$), the VAE objective is a minimization of variational lower bound of the negative log likelihood:

$$L(\alpha, \beta) = \sum_{i=1}^{\tilde{n}} [-\log q_\alpha(\tilde{x}_i) + \gamma \mathrm{KL}(\pi_\beta(z_i|\tilde{x}_i) || q_\alpha(z_i|\tilde{x}_i))] \quad (12)$$

[5] Jianwen Xie, Zilong Zheng, Ping Li. Energy-Based Probability Estimationwith Variational Ancestral Langevin Sampler. 2020. (under review)

Variational MCMC teaching



(a) MCMC teaching (b) fast MCMC teaching (c) variational MCMC teaching

Variational MCMC teaching

In general, the benefits of MCMC teaching are

(1) The latent variable model $q_\alpha(x)$ provides an efficient MCMC for the EBM $p_\theta(x)$.

(2) The EBM $p_\theta(x)$ provides infinite training data for the latent variable model $q_\alpha(x)$.

EBM with variational ancestral Langevin sampler

Algorithm 1 Learning EBM with Variational Ancestral Langevin Sampler

Input: : (1) training images $\{x_i\}_{i=1}^{n}$, (2) numbers of Langevin steps l

Output: : (1) parameters $\{\theta, \alpha, \beta\}$, (2) initial samples $\{\hat{x}_i\}_{i=1}^{\tilde{n}}$, (3) Langevin samples $\{\tilde{x}_i\}_{i=1}^{\tilde{n}}$

1: Let $t \leftarrow 0$, initialize θ, α, and β.
2: **repeat**
3: **Ancestral Langevin Sampling**: For $i = 1, ..., \tilde{n}$, sample $\hat{z}_i \sim \mathrm{N}(0, I_d)$, then generate $\hat{x}_i = g(\hat{z}_i)$, and run l steps of Langevin revision dynamics from $\hat{x}_i$ to obtain $\tilde{x}_i$, each step following Eq. (6)(ii).
4: **Maximum Likelihood Learning**: Treat $\{\tilde{x}\}_i^{\tilde{n}}$ as MCMC examples from $p_\theta(x)$, update θ by Adam with the gradient computed according to Eq. (7).
5: **Variational Auto-Encoding**: Treat $\{\tilde{x}\}_i^{\tilde{n}}$ as training data, update α and β by minimizing VAE objective in Eq. (9) via Adam.
6: Let $t \leftarrow t + 1$
7: **until** $t = T$

Nash equilibrium

Hinton proposed Contrastive Divergence [6] to train RBM (a special EBM). CD runs k steps of MCMC initialized from the training data, instead for Gaussian noise.

Contrastive divergence (CD)

Given an energy-based model $p_\theta(x)$. Let M_θ be the transition kernel of the finite-step MCMC that samples from $p_\theta(x)$.

$$\hat{\theta} = \arg\min_\theta [\mathrm{KL}(p_{\mathrm{data}}(x) \| p_\theta(x)) - \mathrm{KL}(\mathcal{M}_\theta p_{\mathrm{data}}(x) \| p_\theta(x))], \quad (13)$$

If $\mathcal{M}_\theta p_{\mathrm{data}}(x)$ is close to p_θ, then the second divergence is small, and the CD estimate is close to maximum likelihood which minimizes the first divergence.

[6] GE Hinton. Training products of experts by minimizing contrastive divergence. Neural computation, 2002

Nash equilibrium

A Nash equilibrium of the model is a triplet $(\hat{\theta}, \hat{\alpha}, \hat{\beta})$ that satisfies:

$$\hat{\theta} = \arg\min_{\theta} [\mathrm{KL}(p_{\mathrm{data}}(x) \| p_{\theta}(x)) - \mathrm{KL}(\mathcal{M}_{\hat{\theta}} q_{\hat{\alpha}}(x) \| p_{\theta}(x))], \tag{14}$$

$$\hat{\alpha} = \arg\min_{\alpha} [\mathrm{KL}(\mathcal{M}_{\hat{\theta}} q_{\hat{\alpha}}(x) \| q_{\alpha}(x)) + \mathrm{KL}(\pi_{\hat{\beta}}(z|x) \| q_{\alpha}(z|x))], \tag{15}$$

$$\hat{\beta} = \arg\min_{\beta} \mathrm{KL}(\pi_{\beta}(z|x) \| q_{\hat{\alpha}}(z|x)), \tag{16}$$

We show that if $(\hat{\theta}, \hat{\alpha}, \hat{\beta})$ is a Nash equilibrium of the model, then $p_{\hat{\theta}} = q_{\hat{\alpha}} = p_{\mathrm{data}}$.

Understanding the learning dynamics

The proposed framework includes three trainable models, i.e., energy-based model $p_\theta(x)$, inference model $\pi_\beta(z|x)$, and latent variable model $q_\alpha(x|z)$. They, along with the empirical data distribution p_{data} and the Gaussian prior distribution $q(z)$, define three joint distributions over the latent variables z and the data x.

Three joint distributions

(1) Π-distribution: $\Pi(z, x) = p_{\text{data}}(x)\pi_\beta(z|x)$
(2) Q-distribution: $Q(z, x) = q(z)q_\alpha(x|z)$
(3) P-distribution: $P(z, x) = p_\theta(x)\pi_\beta(z|x)$

Understanding the learning dynamics

VAEs learn $\{\alpha, \beta\}$ from training data $p_{\rm data}$, whose objective function is $\min_\beta \min_\alpha \mathrm{KL}(\Pi||Q)$.

The VAE learns to mimic the EBM at each iteration by learning from its generated examples. Thus, given θ_t at iteration t, the VAE objective becomes $\min_\beta \min_\alpha \mathrm{KL}(P_{\theta_t}||Q)$, where we put subscript θ_t in P to indicate that the P distribution is associated with a fixed θ_t.

$$
\begin{aligned}
&\mathrm{KL}(P_{\theta_t}||Q) \\
=&\mathrm{KL}(p_{\theta_t}(x)\pi_\beta(z|x)||q_\alpha(x|z)q(z)) \\
=&\mathrm{KL}(p_{\theta_t}(x)||q_\alpha(x)) + \mathrm{KL}(\pi_\beta(z|x)||q_\alpha(z|x)) \\
=&\mathrm{KL}(\mathcal{M}_{\theta_t}q_{\alpha_t}(x)||q_\alpha(x)) + \mathrm{KL}(\pi_\beta(z|x)||q_\alpha(z|x)) \\
&.
\end{aligned}
\tag{17}
$$

Understanding the learning dynamics

Three joint distributions

(1) Π-distribution: $\Pi(z, x) = p_{\mathrm{data}}(x)\pi_\beta(z|x)$
(2) Q-distribution: $Q(z, x) = q(z)q_\alpha(x|z)$
(3) P-distribution: $P(z, x) = p_\theta(x)\pi_\beta(z|x)$

The joint minimization in VAE can be interpreted as alternating projection between P_{θ_t} and Q, where π_β and q_α run toward each other and eventually converge at the intersection.

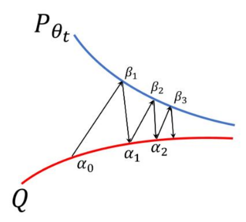


Figure 1: Training variational auto-encoder (VAE) by alternating projection.

Understanding the learning dynamics

Three joint distributions

(1) Π-distribution: $\Pi(z, x) = p_{\mathrm{data}}(x)\pi_\beta(z|x)$
(2) Q-distribution: $Q(z, x) = q(z)q_\alpha(x|z)$
(3) P-distribution: $P(z, x) = p_\theta(x)\pi_\beta(z|x)$

With the examples generated by the ancestral Langevin sampler, the objective function of training the EBM is $\min_\theta \mathrm{KL}(\Pi||P)$, i.e., $\min_\theta \mathrm{KL}(p_{\mathrm{data}}||p_\theta)$.

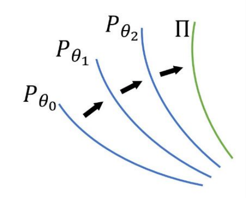


Figure 2: Energy-based learning via distribution shifting

Understanding the learning dynamics

Three joint distributions

(1) Π-distribution: $\Pi(z, x) = p_{\mathrm{data}}(x)\pi_\beta(z|x)$
(2) Q-distribution: $Q(z, x) = q(z)q_\alpha(x|z)$
(3) P-distribution: $P(z, x) = p_\theta(x)\pi_\beta(z|x)$

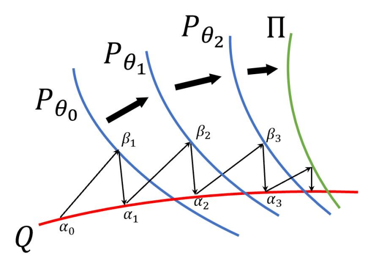


Figure 3: Motional alternating projection

Understanding the learning dynamics

Three joint distributions

(1) Π-distribution: $\Pi(z, x) = p_{\mathrm{data}}(x)\pi_\beta(z|x)$
(2) Q-distribution: $Q(z, x) = q(z)q_\alpha(x|z)$
(3) P-distribution: $P(z, x) = p_\theta(x)\pi_\beta(z|x)$

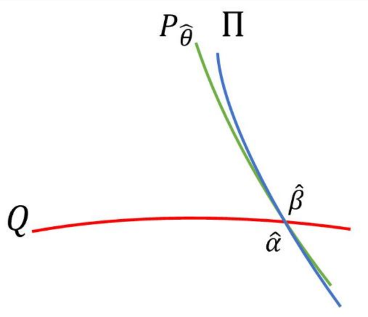


Figure 4: Convergent point of the motional alternating projection

Experiments: Image Generation

Figure 5: Generated Samples by the model learned on MNIST, Fashion-MNIST and Cifar-10 datasets.

Experiments: Image Generation

Model	IS
PixelCNN (Van den Oord et al. 2016)	4.60
PixelIQN (Ostrovski, Dabney, and Munos 2018)	5.29
EBM (Du and Mordatch 2019)	6.02
DCGAN (Radford, Metz, and Chintala 2015)	6.40
WGAN+GP (Gulrajani et al. 2017)	6.50
CoopNets (Xie et al. 2018a)	6.55
VALS (Ours)	**6.65**

Figure 6: Quantitative evaluation of Inception score and FID score on CIFAR-10 dataset

Experiments: Image Generation

(a) Interpolation by the latent variable model

(b) Langevin revision by a learned model

Experiments: Conditional Image Generation

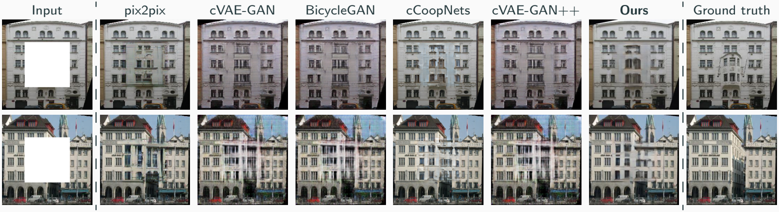


Figure 7: Example results of image completion on facades testing dataset.

Table 1: Comparison with the baselines for image inpainting

	CMP Facades		**Paris StreetView**	
Method	PSNR	SSIM	PSNR	SSIM
pix2pix	19.34	0.74	15.17	0.75
cVAE-GAN	19.43	0.68	16.12	0.72
cVAE-GAN++	19.14	0.64	16.03	0.69
BicycleGAN	19.07	0.64	16.00	0.68
cCoopNets	20.47	0.77	21.17	0.79
VALS (Ours)	**21.62**	**0.78**	**22.61**	**0.79**

Conclusion

- We present a new framework to train EBM jointly with a VAE via MCMC teaching.
- We provide a new strategy, variational MCMC teaching, to train latent variable model (generator).
- We naturally unify the maximum likelihood learning (MLE), variational inference and MCMC teaching in a single framework.
- We demonstrate empirical results on both unconditional and conditional image models.