Learning Energy-Based Model with Variational Auto-Encoder as Amortized Sampler

Jianwen Xie, Zilong Zheng, Ping Li

Cognitive Computing Lab, Baidu Research, Bellevue, USA

Introduction and motivation

1. A persisting challenge in training energy-based models (EBMs) is the calculation of the intractable normalizing constant, which typically requires Markov chain Monte Carlo (MCMC).
2. However, the MCMC is computationally expensive or even impractical.
3. To tackle the challenge, this paper learns a variational auto-encoder (VAE) as an amortized sampler for efficient training of EBMs.

Contribution

1. We propose to learn a variational auto-encoder (VAE) to initialize the finite-step MCMC, such as Langevin dynamics, for efficient amortized sampling of the EBM.
2. We naturally unify the maximum likelihood learning, variational inference, and MCMC teaching in a single framework.
3. We provide an information geometric understanding of the proposed joint training algorithm. It can be interpreted as a dynamic alternating projection.
4. We provide strong empirical results on unconditional image modeling and conditional predictive learning to validate the proposed method.

Energy-based model and "analysis by synthesis"

(1) Energy-based Model
Let x be an image, $U_\theta(x)$ be an energy function where θ is trainable parameters, an EBM is defined as a probability density:

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

where $Z(\theta) = \int \exp[-U_\theta(x)]dx$ is an analytically intractable normalizing constant. Following the EBM introduced by Xie et al.(2016) [a], we can parameterize $U_\theta(x)$ by a bottom-up ConvNet with weights θ and scalar output.

(2) Analysis by synthesis
Suppose we have a training set $\mathcal{D} = \{x_i, i = 1, ..., n\}$ and we assume each datapoint is sampled from an unknown distribution $p_{data}(x)$. We train θ by maximum likelihood. The gradient is computed by

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

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). This will lead to an "analysis by synthesis" algorithm that iterates a synthesis step for image sampling and an analysis step for parameter learning.

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

A gradient-based MCMC: Langevin dynamics

(1) Langevin dynamics
Given the current energy function $U_\theta(x)$, the Langevin Dynamics iterates

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

where t indexes the time step, δ is the step size, and the initial state $\tilde{x}_0$ follows a uniform distribution.

(2) Challenges

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

Variational auto-encoder as amortized sampler

(1) Ancestral Langevin Sampling
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 iterative Langevin sampler. We draw a sample by first (i) sampling an initial example $\hat{x}$ via ancestral sampling, and then (ii) revising $\hat{x}$ with a finite-step Langevin update, that is

$$\text{(i)}\hat{x} = g_\alpha(\hat{z}), \hat{z} \sim \mathcal{N}(0, I_d),$$

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

The goal of $g_\alpha(z)$ is to provide a good starting point for MCMC sampling, i.e., mimic the the distribution of $p_\theta(x)$.

(2) Update EBM $p_\theta(x)$ with amortized sampling
With $\{\tilde{x}_i\}_{i=1}^{\tilde{n}} \sim p_\theta(x)$ via ancestral Langevin sampling, we can compute the gradient for θ by

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

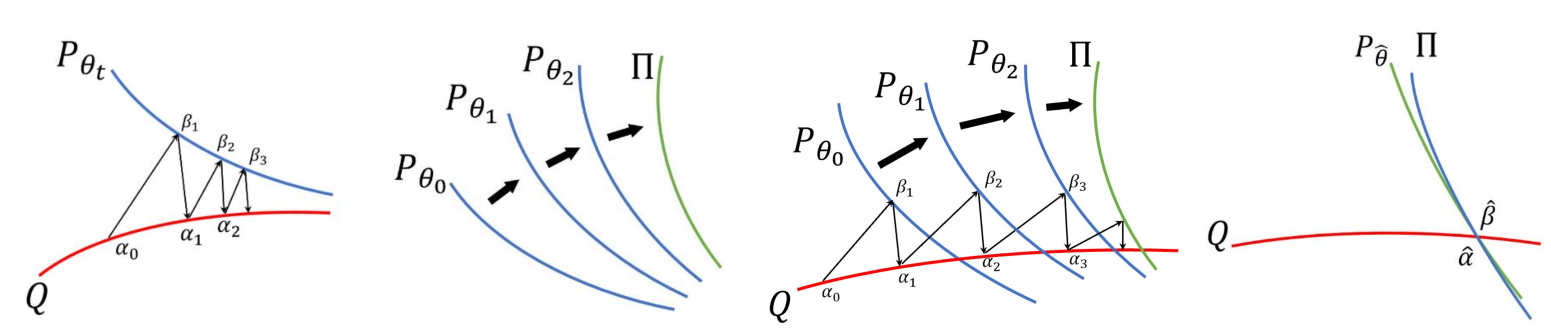
and then update θ by Adam optimizer.

(3) Update latent variable model $q_\alpha(x)$ by variational MCMC teaching
In our paper, we want to learn $q_\alpha(x)$ in the context of 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. The VAE objective is a minimization of variational lower bound of the negative log likelihood:

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

Information geometric understanding

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



(a) (b) (c) (d)

(a) Variational learning as alternating projection
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.
(b) Energy-based learning as manifold shifting
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., $\min_\theta \mathrm{KL}(p_{\mathrm{data}}||p_\theta)$. P_{θ_0} runs toward Π and seeks to match it.
(c) Integrating energy-based learning and variational learning as dynamic alternating projection
Our model can be interpreted as a dynamic alternating projection between Q and P, where Q is static but P is changeable and keeps shifting toward Π.
(d) Convergent point of the dynamic alternating projection
Triplet $(\hat{\theta}, \hat{\alpha}, \hat{\beta})$ is the Nash equilibrium (optimal solution) of the learning algorithm.

Experiment 1: Image generation

Figure 1: Generated samples by the models learned on MNIST, Fashion-MNIST and CIFAR-10.

Table 1: Quantitative evaluation of Inception score and FID score on CIFAR-10 dataset

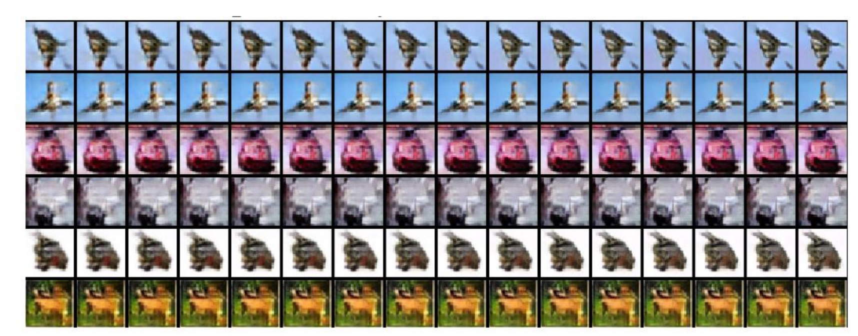
Model	IS	FID
PixelCNN (Van den Oord et al. 2016)	4.60	65.93
PixelIQN (Ostrovski, Dabney, and Munos 2018)	5.29	49.46
EBM (Du and Mordatch 2019)	6.02	40.58
DCGAN (Radford, Metz, and Chintala 2016)	6.40	37.11
WGAN+GP (Gulrajani et al. 2017)	6.50	36.4
CoopNets (Xie et al. 2018a)	6.55	36.4
Ours	**6.65**	**36.2**

Experiment 2: Model analysis

We check whether the latent variable model learns a meaningful latent space in the proposed learning scheme by demonstrating interpolation between generated examples.

(a) Interpolation by the latent variable model

We also check the gap between the EBM and the latent variable model once they are leaned. Visualization of ancestral Langevin dynamics when the model converges. For each row, the leftmost image is the synthesized output by the ancestral sampling. The rest image sequence displays the synthesized images revised at different Langevin steps.



(b) Langevin revision by a learned model

Experiment 3: Image recovery

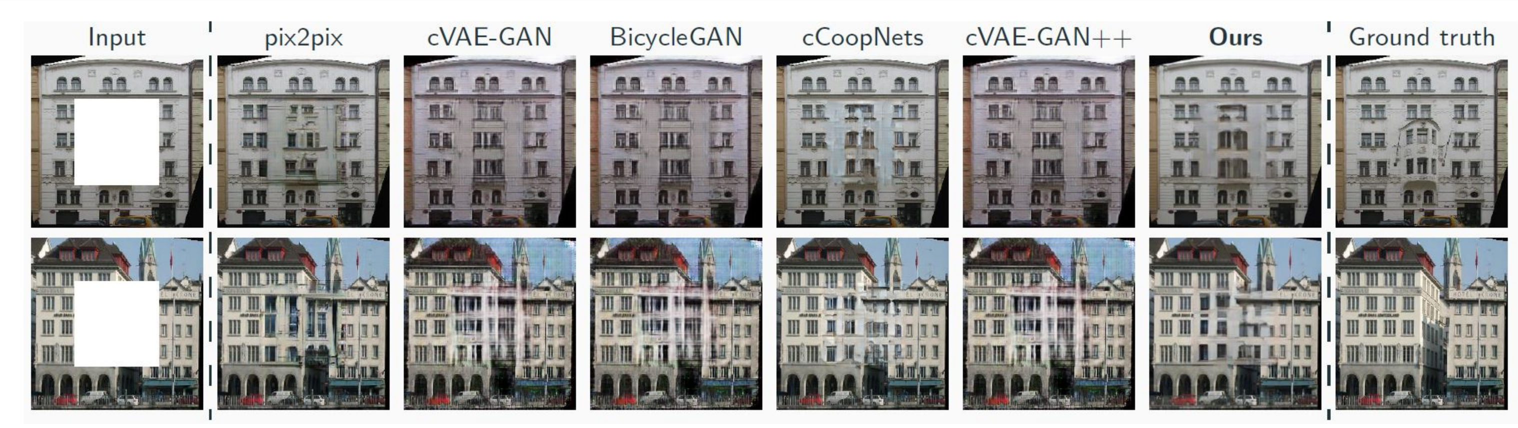


Figure 2: Example results of image recovery on facades testing dataset.

Table 2: Comparison with the baselines for image inpainting

Method	CMP Facades PSNR	CMP Facades SSIM	Paris StreetView PSNR	Paris StreetView SSIM
pix2pix (Isola et al. 2017)	19.34	0.74	15.17	0.75
cVAE-GAN (Zhu et al. 2017)	19.43	0.68	16.12	0.72
cVAE-GAN++ (Zhu et al. 2017)	19.14	0.64	16.03	0.69
BicycleGAN (Zhu et al. 2017)	19.07	0.64	16.00	0.68
cCoopNets (Xie et al. 2018a)	20.47	0.77	21.17	0.79
Ours	**21.62**	**0.78**	**22.61**	**0.79**