Learning Energy-based Model via Dual-MCMC Teaching

Jiali Cui Tian Han

Department of Computer Science, Stevens Institute of Technology

October 20, 2023

Energy-based Model (EBM)

- The EBM can be specified with the probabilistic density:

$$\pi_\alpha(\mathbf{x}) = \frac{1}{\mathrm{Z}(\alpha)} \exp\left[f_\alpha(\mathbf{x})\right]$$

- The EBM can be learned via Maximum Likelihood Estimation (MLE):

$$\max_\alpha L_\pi(\alpha) = \frac{1}{n} \sum_{i=1}^{n} \log \pi_\alpha(\mathbf{x}^{(i)})$$

- The learning gradient can be computed as

$$\mathbb{E}_{p_{\text{data}}(\mathbf{x})}\left[\frac{\partial}{\partial \alpha} f_\alpha(\mathbf{x})\right] - \mathbb{E}_{\pi_\alpha(\mathbf{x})}\left[\frac{\partial}{\partial \alpha} f_\alpha(\mathbf{x})\right]$$

EBM Sampling

- The MLE learning requires sampling from EBM, and it can be achieved by MCMC sampling, such as Langevin dynamics:

$$\mathbf{x}_{\tau+1} = \mathbf{x}_\tau + s\frac{\partial}{\partial \mathbf{x}_\tau}\log \pi_\alpha(\mathbf{x}_\tau) + \sqrt{2s}U_\tau$$

where τ indexes the time step, s is the step size and $U_\tau \sim \mathcal{N}(0, I_D)$.

- However, noise-initialized Langevin dynamics can be extremely inefficient and ineffective as they usually take a long time to converge between different modes and are also non-stable in practice

Generator Model

- The generator model can be specified as joint density:

$$p_\theta(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p_\theta(\mathbf{x}|\mathbf{z})$$

- The generator model can be learned via MLE:

$$\max_\theta L_p(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(\mathbf{x}^{(i)}), \text{ where } p_\theta(\mathbf{x}) = \int_{\mathbf{z}} p_\theta(\mathbf{x}, \mathbf{z})d\mathbf{z}$$

- The learning gradient can be computed as

$$\mathbb{E}_{p_{\text{data}}(\mathbf{x})p_\theta(\mathbf{z}|\mathbf{x})}\left[\frac{\partial}{\partial\theta} \log p_\theta(\mathbf{x}, \mathbf{z})\right]$$

Generator Posterior Sampling

- The MLE learning requires sampling from generator posterior, and it can be achieved by latent space MCMC sampling:

$$\mathbf{z}_{\tau+1} = \mathbf{z}_\tau + s \frac{\partial}{\partial \mathbf{z}_\tau} \log p_\theta(\mathbf{z}_\tau|\mathbf{x}) + \sqrt{2s} U_\tau$$

- However, noise-initialized Langevin dynamics can be ineffective in traversing the latent space and hard to mix.

Inference Model

Motivation: For the MLE learning of the EBM, MCMC sampling can be initialized through the complementary generator model, while for the MLE learning of the generator model, the latent MCMC sampling can be initialized by the complementary **inference model**.

- Inference model is assumed to be Gaussian:

$$q_\phi(\mathbf{z}|\mathbf{x}) \sim \mathcal{N}(\mu_\phi(\mathbf{x}), V_\phi(\mathbf{x}))$$

where $\mu_\phi(\mathbf{x})$ is the mean d-dimensional mean vector and $V_\phi(\mathbf{x})$ is the d-dimensional diagonal covariance matrix.

Joint Densities & Dual-MCMC Teaching

- With the EBM $\pi_\alpha(\mathbf{x})$, generator $p_\theta(\mathbf{x})$ and the inference model $q_\phi(\mathbf{z}|\mathbf{x})$, we can naturally specify the three densities on joint space $(\mathbf{x}, \mathbf{z})$, i.e.,

$$P_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z}),$$
$$\Pi_{\alpha,\phi}(\mathbf{x}, \mathbf{z}) = \pi_\alpha(\mathbf{x})q_\phi(\mathbf{z}|\mathbf{x}),$$
$$Q_\phi(\mathbf{x}, \mathbf{z}) = p_{\text{data}}(\mathbf{x})q_\phi(\mathbf{z}|\mathbf{x})$$

- We introduce two joint distributions that incorporate MCMC sampling as revision processes

$$\tilde{P}_{\theta,\alpha}(\mathbf{x}, \mathbf{z}) = \mathcal{T}_\alpha^{\mathbf{x}} p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z}) \quad \tilde{Q}_{\phi,\theta}(\mathbf{x}, \mathbf{z}) = p_{\text{data}}(\mathbf{x})\mathcal{T}_\theta^{\mathbf{z}} q_\phi(\mathbf{z}|\mathbf{x})$$

Learning via Dual-MCMC Teaching

- For EBM $\pi_\alpha(\mathbf{x})$ learning,

$$\max_\alpha \tilde{L}_\pi(\alpha) = \min_\alpha \mathrm{KL}(\tilde{Q}_{\phi_t,\theta_t}(\mathbf{x},\mathbf{z})\|\Pi_{\alpha,\phi}(\mathbf{x},\mathbf{z})) - \mathrm{KL}(\tilde{P}_{\theta_t,\alpha_t}(\mathbf{x},\mathbf{z})\|\Pi_{\alpha,\phi}(\mathbf{x},\mathbf{z}))$$

- For generator model $p_\theta(\mathbf{x})$ learning,

$$\max_\theta \tilde{L}_p(\theta) = \min_\theta \mathrm{KL}(\tilde{Q}_{\phi_t,\theta_t}(\mathbf{x},\mathbf{z})\|P_\theta(\mathbf{x},\mathbf{z})) + \mathrm{KL}(\tilde{P}_{\theta_t,\alpha_t}(\mathbf{x},\mathbf{z})\|P_\theta(\mathbf{x},\mathbf{z}))$$

- For inference model $q_\phi(\mathbf{z}|\mathbf{x})$ learning,

$$\min_\phi D_q(\phi) = \min_\phi \mathrm{KL}(\tilde{Q}_{\phi_t,\theta_t}(\mathbf{x},\mathbf{z})\|Q_\phi(\mathbf{x},\mathbf{z})) + \mathrm{KL}(\tilde{P}_{\theta_t,\alpha_t}(\mathbf{x},\mathbf{z})\|\Pi_{\alpha,\phi}(\mathbf{x},\mathbf{z}))$$

Image Modelling

Table: FID and IS on CIFAR-10 and CelebA-64.

Methods	CIFAR-10		CelebA-64
	IS (↑)	FID (↓)	FID (↓)
Ours	**8.55**	**9.26**	**5.15**
Cooperative EBM	6.55	33.61	16.65
Amortized EBM	6.65	-	-
Divergence Triangle	7.23	30.10	18.21
No MCMC EBM	-	27.5	-
Short-run EBM	6.21	-	23.02
IGEBM	6.78	38.2	-
ImprovedCD EBM	7.85	25.1	-
Diffusion EBM	8.30	9.58	5.98
VAEBM	8.43	12.19	5.31
NCP-VAE	-	24.08	5.25
SNGAN	8.22	21.7	6.1
StyleGANv2 w/o ADA	8.99	9.9	2.32
NCSN	8.87	25.32	25.30
DDPM	9.46	3.17	3.93

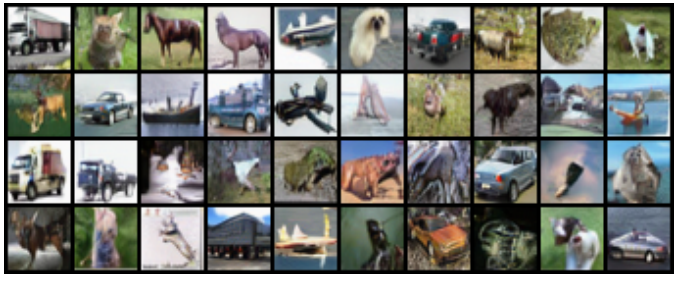

Figure: Visualization of Latent Space.

MCMC-Revision

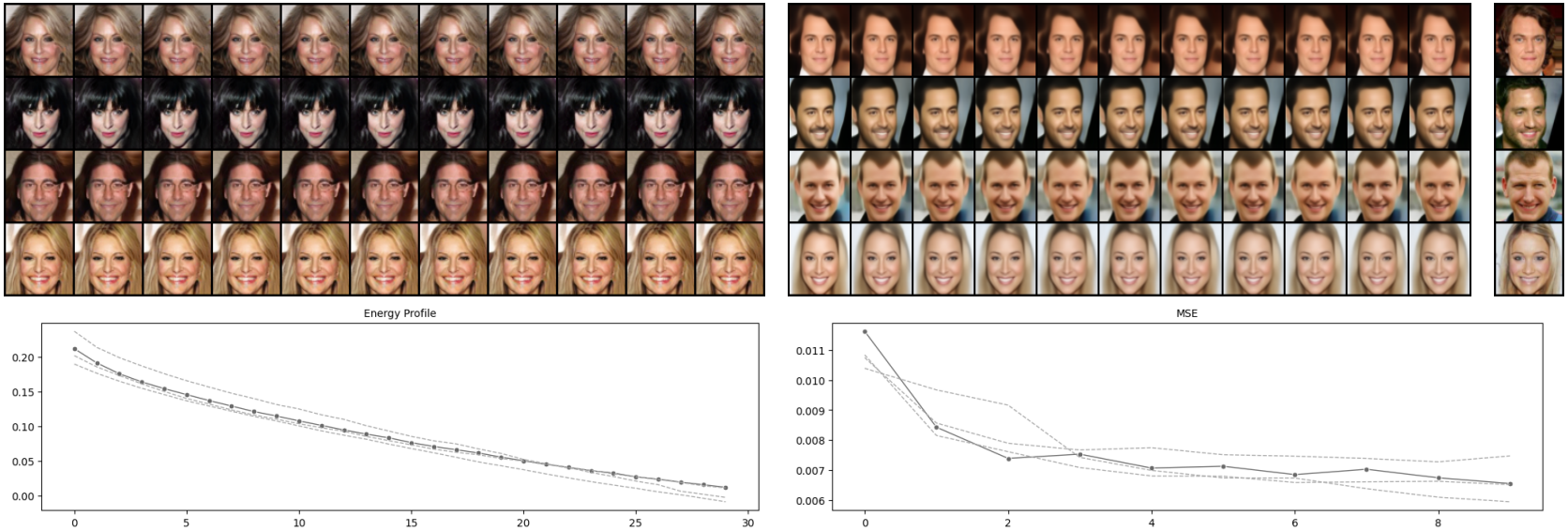

Figure: **Left top:** MCMC revision on $\mathbf{x}$. The leftmost images are sampled from the generator model, and the rightmost images are at the final step of the EBM-guided MCMC sampling. **Left bottom:** Energy profile over steps. **Right top:** MCMC revision on $\mathbf{z}$. The leftmost images are reconstructed by latent codes inferred from the inference model, and the rightmost images are reconstructed by latent codes at the final step of the generator-guided MCMC inference. **Right bottom:** Mean Squared Error (MSE) over steps.