Efficient Training of Energy-Based Models Using Jarzynski Equality

Davide Carbone (Politecnico di Torino), Mengjian Hua (New York University),
Simon Coste (University of Paris - P7) and Eric Vanden-Eijnden (New York University)

Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS 2023)

EBM I

- **Definition**: Energy-Based models (EBMs) are families of parametrized pdf

$$\rho_\theta(x) = Z_\theta^{-1} e^{-U_\theta(x)}; \qquad Z_\theta = \int_{\mathbb{R}^d} e^{-U_\theta(x)} dx$$

where $U_\theta : \mathbb{R}^d \to [0, \infty)$ is the energy function. The target density $\rho_*(x)$ we would like to fit is known just through samples $\{x_*^i\}_{i=1}^n \sim \rho_*(x)$.

EBM I

- **Definition**: Energy-Based models (EBMs) are families of parametrized pdf

$$\rho_\theta(x) = Z_\theta^{-1} e^{-U_\theta(x)}; \qquad Z_\theta = \int_{\mathbb{R}^d} e^{-U_\theta(x)} dx$$

  where $U_\theta : \mathbb{R}^d \to [0, \infty)$ is the energy function. The target density $\rho_*(x)$ we would like to fit is known just through samples $\{x_*^i\}_{i=1}^n \sim \rho_*(x)$.

- **Training**: gradient descent over cross entropy (i.e. over KL divergence up to a constant)

$$\dot{\theta}(t) = -\partial_\theta H(\rho_\theta, \rho_*) = \underbrace{-\mathbb{E}_*[\partial_\theta U_\theta]}_{\text{from data}} +$$

EBM I

- **Definition**: Energy-Based models (EBMs) are families of parametrized pdf

$$\rho_\theta(x) = Z_\theta^{-1} e^{-U_\theta(x)}; \qquad Z_\theta = \int_{\mathbb{R}^d} e^{-U_\theta(x)} dx$$

  where $U_\theta : \mathbb{R}^d \to [0, \infty)$ is the energy function. The target density $\rho_*(x)$ we would like to fit is known just through samples $\{x_*^i\}_{i=1}^n \sim \rho_*(x)$.

- **Training**: gradient descent over cross entropy (i.e. over KL divergence up to a constant)

$$\dot{\theta}(t) = -\partial_\theta H(\rho_\theta, \rho_*) = \underbrace{-\mathbb{E}_*[\partial_\theta U_\theta]}_{\text{from data}} + \underbrace{\mathbb{E}_\theta[\partial_\theta U_\theta]}_{\text{samples from } \rho_\theta}$$

EBM I

- **Definition**: Energy-Based models (EBMs) are families of parametrized pdf

$$\rho_\theta(x) = Z_\theta^{-1} e^{-U_\theta(x)}; \qquad Z_\theta = \int_{\mathbb{R}^d} e^{-U_\theta(x)} dx$$

  where $U_\theta : \mathbb{R}^d \to [0, \infty)$ is the energy function. The target density $\rho_*(x)$ we would like to fit is known just through samples $\{x_*^i\}_{i=1}^n \sim \rho_*(x)$.

- **Training**: gradient descent over cross entropy (i.e. over KL divergence up to a constant)

$$\dot{\theta}(t) = -\partial_\theta H(\rho_\theta, \rho_*) = \underbrace{-\mathbb{E}_*[\partial_\theta U_\theta]}_{\text{from data}} + \underbrace{\mathbb{E}_\theta[\partial_\theta U_\theta]}_{\text{samples from } \rho_\theta}$$

- **Possible solution**: generate samples from ρ_θ to compute $E_\theta[\partial_\theta U_\theta]$, using for instance ULA, MALA, Gibbs sampling, etc.

- **Mixing**: at θ fixed, ULA is the Markov process

$$X_{k+1} = X_k - h\nabla U_\theta(X_k) + \sqrt{2h}\xi_k, \qquad X_0 \sim \rho_0$$

for $k \geq 0$, $h > 0$ and $\{\xi_k\}_{k\in\mathbb{N}_0}$ are independent $\mathcal{N}(0_d, I_d)$.

EBM II

- **Mixing**: at θ fixed, ULA is the Markov process

$$X_{k+1} = X_k - h\nabla U_\theta(X_k) + \sqrt{2h}\xi_k, \qquad X_0 \sim \rho_0$$

for $k \geq 0$, $h > 0$ and $\{\xi_k\}_{k\in\mathbb{N}_0}$ are independent $\mathcal{N}(0_d, I_d)$. We have $X_\infty \sim \rho_\theta$

EBM II

- **Mixing**: at θ fixed, ULA is the Markov process

$$X_{k+1} = X_k - h\nabla U_\theta(X_k) + \sqrt{2h}\xi_k, \qquad X_0 \sim \rho_0$$

for $k \geq 0$, $h > 0$ and $\{\xi_k\}_{k\in\mathbb{N}_0}$ are independent $\mathcal{N}(0_d, I_d)$. We have $X_\infty \sim \rho_\theta$ but given $X_k \sim \rho_k(x)$

$$\mathbb{E}_\theta[\partial_\theta U_\theta] \neq \int \partial_\theta U_\theta(x)\rho_k(x)\,dx$$

for any $k < \infty$. Even less controlled along EBM training since $\theta = \theta_k$ depends on time.

EBM II

- **Mixing**: at θ fixed, ULA is the Markov process

$$X_{k+1} = X_k - h\nabla U_\theta(X_k) + \sqrt{2h}\xi_k, \qquad X_0 \sim \rho_0$$

for $k \geq 0$, $h > 0$ and $\{\xi_k\}_{k\in\mathbb{N}_0}$ are independent $\mathcal{N}(0_d, I_d)$. We have $X_\infty \sim \rho_\theta$ but given $X_k \sim \rho_k(x)$

$$\mathbb{E}_\theta[\partial_\theta U_\theta] \neq \int \partial_\theta U_\theta(x)\rho_k(x)\,dx$$

for any $k < \infty$. Even less controlled along EBM training since $\theta = \theta_k$ depends on time.

- **State of the Art**: Constrastive Divergence ($\rho_0 = \rho_*$ with reinitialization of the chain at ρ_*) and Persistent Contrastive Divergence ($\rho_0 = \rho_*$). CD effectively performs GD on Fisher divergence [**Domingo-Enrich et al., 2021**].

EBM III

- **Main result**: given the dicrete-time dynamical system

$$\begin{cases} X_{k+1} = X_k - h\nabla U_{\theta_k}(X_k) + \sqrt{2h}\,\xi_k, & X_0 \sim \rho_{\theta_0}, \\ A_{k+1} = A_k - \alpha_{k+1}(X_{k+1}, X_k) + \alpha_k(X_k, X_{k+1}), & A_0 = 0, \end{cases}$$

with

$$\alpha_k(x,y) = U_{\theta_k}(x) + \tfrac{1}{2}(y-x)\cdot\nabla U_{\theta_k}(x) + \tfrac{1}{4}h|\nabla U_{\theta_k}(x)|^2$$

- **Main result**: given the dicrete-time dynamical system

$$\begin{cases} X_{k+1} = X_k - h\nabla U_{\theta_k}(X_k) + \sqrt{2h}\xi_k, & X_0 \sim \rho_{\theta_0}, \\ A_{k+1} = A_k - \alpha_{k+1}(X_{k+1}, X_k) + \alpha_k(X_k, X_{k+1}), & A_0 = 0, \end{cases}$$

with

$$\alpha_k(x,y) = U_{\theta_k}(x) + \tfrac{1}{2}(y-x)\cdot\nabla U_{\theta_k}(x) + \tfrac{1}{4}h|\nabla U_{\theta_k}(x)|^2$$

for all $k \in \mathbb{N}_0$, the following equalities hold

$$\mathbb{E}_{\theta_k}[\partial_\theta U_{\theta_k}] = \frac{\mathbb{E}[\partial_\theta U_{\theta_k}(X_k)e^{A_k}]}{\mathbb{E}[e^{A_k}]}, \qquad Z_{\theta_k} = Z_{\theta_0}\mathbb{E}\left[e^{A_k}\right]$$

where $\mathbb{E}[\cdot]$ is the expectation w.r.t. the law of the joint process $(X_k, A_k) \in \mathbb{R}^d \times \mathbb{R}$

- **Main result**: given the dicrete-time dynamical system

$$\begin{cases} X_{k+1} = X_k - h\nabla U_{\theta_k}(X_k) + \sqrt{2h}\xi_k, & X_0 \sim \rho_{\theta_0}, \\ A_{k+1} = A_k - \alpha_{k+1}(X_{k+1}, X_k) + \alpha_k(X_k, X_{k+1}), & A_0 = 0, \end{cases}$$

with

$$\alpha_k(x,y) = U_{\theta_k}(x) + \tfrac{1}{2}(y-x)\cdot\nabla U_{\theta_k}(x) + \tfrac{1}{4}h|\nabla U_{\theta_k}(x)|^2$$

for all $k \in \mathbb{N}_0$, the following equalities hold

$$\mathbb{E}_{\theta_k}[\partial_\theta U_{\theta_k}] = \frac{\mathbb{E}[\partial_\theta U_{\theta_k}(X_k)e^{A_k}]}{\mathbb{E}[e^{A_k}]}, \qquad Z_{\theta_k} = Z_{\theta_0}\mathbb{E}\left[e^{A_k}\right]$$

where $\mathbb{E}[\cdot]$ is the expectation w.r.t. the law of the joint process $(X_k, A_k) \in \mathbb{R}^d \times \mathbb{R}$

EBM III

- **Main result**: given the dicrete-time dynamical system

$$\begin{cases} X_{k+1} = X_k - h\nabla U_{\theta_k}(X_k) + \sqrt{2h}\,\xi_k, & X_0 \sim \rho_{\theta_0}, \\ A_{k+1} = A_k - \alpha_{k+1}(X_{k+1}, X_k) + \alpha_k(X_k, X_{k+1}), & A_0 = 0, \end{cases}$$

with

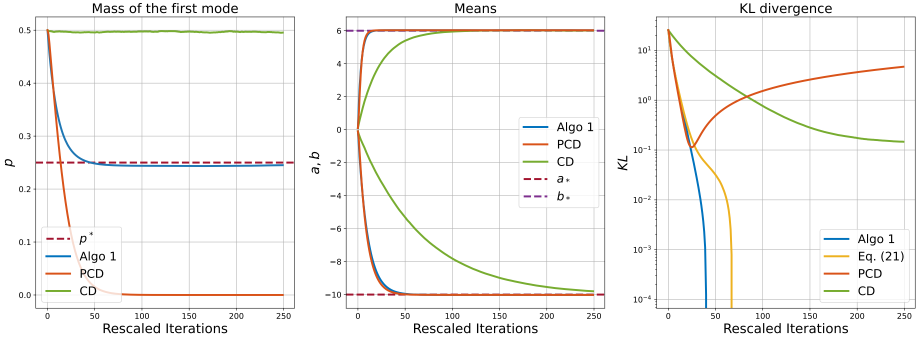
$$\alpha_k(x,y) = U_{\theta_k}(x) + \tfrac{1}{2}(y-x)\cdot\nabla U_{\theta_k}(x) + \tfrac{1}{4}h|\nabla U_{\theta_k}(x)|^2$$

for all $k \in \mathbb{N}_0$, the following equalities hold

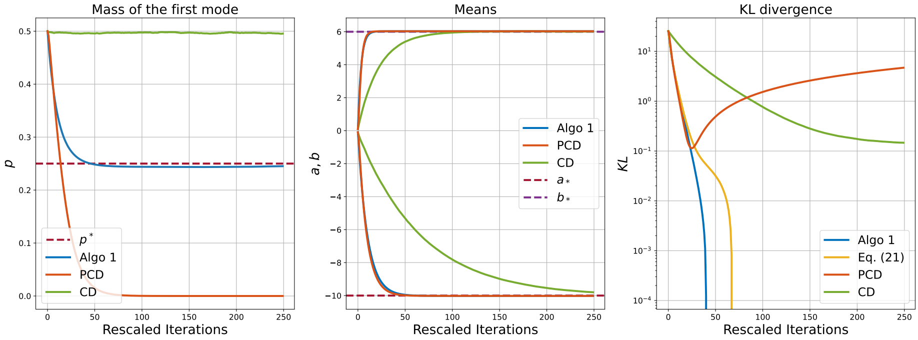
$$\mathbb{E}_{\theta_k}[\partial_\theta U_{\theta_k}] = \frac{\mathbb{E}[\partial_\theta U_{\theta_k}(X_k)e^{A_k}]}{\mathbb{E}[e^{A_k}]}, \qquad \underbrace{\boldsymbol{Z_{\theta_k} = Z_{\theta_0}\mathbb{E}\left[e^{A_k}\right]}}_{\textbf{Jarzynski identity}}$$

where $\mathbb{E}[\cdot]$ is the expectation w.r.t. the law of the joint process $(X_k, A_k) \in \mathbb{R}^d \times \mathbb{R}$

Numerical Experiments I



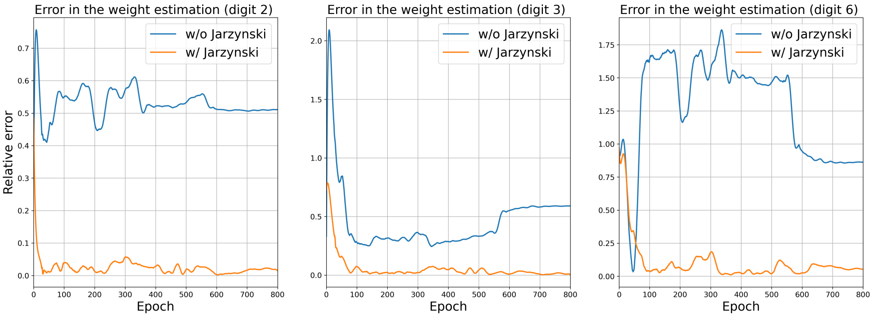
- **Gaussian Mixture**: Algo 1 is our proposal and Eq (21) is the estimation of KL using A_k. PCD and CD does not fit the right relative mass. PCD shows **mode collapse**.



- **Gaussian Mixture**: Algo 1 is our proposal and Eq (21) is the estimation of KL using A_k. PCD and CD does not fit the right relative mass. PCD shows **mode collapse**. CD is performing GD on Fisher divergence, so it is **insensitive to mass imbalance**.

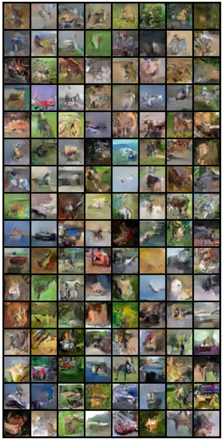
Numerical Experiments II

- **Neural network**: for real datasets like MNIST and CIFAR-10, we use a neural architecture to model the potential
- **MNIST**: we prune the dataset to three digits (2, 3 and 6) in order to stress multimodality and we imbalance the relative number of examples.
- **Jarzynski correction**: we recover the relative mass of the modes

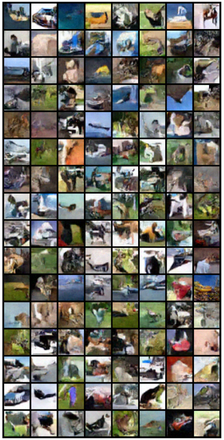


- **CIFAR-10**: for a more complicate dataset, we tried to compare with (almost) state of the art using architectures already present in literature (Nijkamp et al. 2019).

Method	*FID*	*Inception Score (IS)*
PCD with mini-batches	38.25	5.96
PCD with mini-batches and data augmentation	36.43	6.54
Algorithm 4 with multinomial resampling	**32.18**	**6.88**
Algorithm 4 with systematic resampling	**30.24**	**6.97**



Generated CIFAR-10 samples with our approach



Generated CIFAR-10 samples with PCD

Conclusions

Problem

Sampling from a Boltzmann-Gibbs pdf is hard, hence training an EBM using cross-entropy is affected by uncontrolled approximations.

Conclusions

Problem

Sampling from a Boltzmann-Gibbs pdf is hard, hence training an EBM using cross-entropy is affected by uncontrolled approximations.

Solution

Solution: our proposal allows to **exactly perform GD** on cross-entropy. It requires **negligible extra computational cost** and it can be used to substitute any sampling routine (ULA, MALA or others) commonly used in EBM training.