Energy-Based Models

and how to train them

Simon Coste
joint work with Davide Carbone, Mengjian Hua, Eric Vanden-Eijnden

https://arxiv.org/abs/2305.19414

June 9, 2023

Generative Modelling and EBMs

$x_*^1, \ldots, x_*^n$: training samples from an unknown distribution ρ_* ("target")

The two goals of generative modelling:

1. Generate 'new' samples from ρ_* (direct problem)
2. Find a good, interpretable estimator for ρ_* (inverse problem)

EBMs, GANs, VAEs, Normalizing Flows, Neural ODEs, Diffusions, Flow matching...

EBMs

$U_\theta : \mathbb{R}^d \to \mathbb{R}_+ =$ parametrized family of functions ("model energies")

Definition of the model densities:

$$\rho_\theta(x) = \frac{e^{-U_\theta(x)}}{Z_\theta} \qquad\qquad Z_\theta = \int e^{-U_\theta(x)} dx.$$

Which θ_* achieves the best 'fit' between ρ_θ and ρ_*?

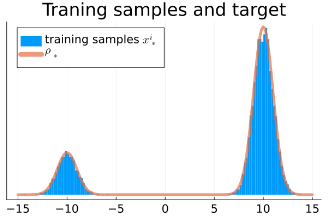
Toy model: Gaussian mixtures

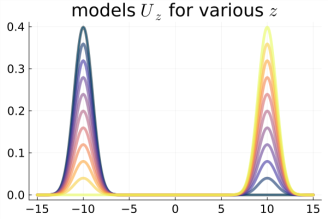
Model: all gaussian mixtures with modes $a = -10, b = 10$:

$$U_z(x) = -\log\left(e^{-|x-a|^2/2} + e^{-z}e^{-|x-b|^2/2}\right)$$

$$Z_z = (1 + e^{-z})\sqrt{2\pi}$$

$$\rho_z(x) = \frac{e^{-|x-a|^2/2} + e^{-z}e^{-|x-b|^2/2}}{(1 + e^{-z})\sqrt{2\pi}}$$





Target: $\rho_* = \rho_{z_*}$ for some z_* with $q_* = \frac{e^{-z_*}}{1+e^{-z_*}} \approx 0.8$.

Training procedures

Score Matching

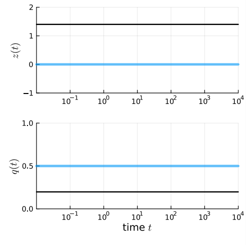
θ_* minimizes the Stein divergence $SM(\theta) = \mathbb{E}_*[|\nabla \log \rho_\theta - \nabla \log \rho_*|^2]$.

Gradient flow

$$\dot{\theta}(t) = -\partial_\theta \mathbb{E}_*[|\nabla \log \rho_{\theta(t)} - \nabla \log \rho_*|^2]$$

Pros: efficiency ([Hyvarinen 2005], [Vincent 2009])

Cons: in the context of high energy barriers, SM cannot learn the relative masses of the energy wells.



Proof of failure.
For any z,

$$\nabla \log \rho_z(x) = \frac{(x-a)e^{-(x-a)^2/2} + e^{-z}(x-b)e^{-(x-b)^2/2}}{e^{-(x-a)^2/2} + e^{-z}e^{-(x-b)^2/2}}$$
$$\approx (x-a)1_{x \text{ close to } a} + (x-b)1_{x \text{ close to } b}$$

$\nabla \log \rho_z(x)$ does not depend on z, hence $\partial_z SM(z) = 0$: this leads to the "no learning" phenomenon,

$$\dot{z}(t) \approx 0$$

□

Gradient ascent on Energy-Based Models

$\theta*$ maximizes the log-likelihood $L(\theta) = \mathbb{E}_*[\log \rho_\theta] = -\mathbb{E}_*[U_\theta + \log Z_\theta]$.

Gradient flow: $\dot{\theta}_t = \partial_\theta L(\theta_t) = -\partial_\theta \log Z_\theta - \mathbb{E}_*[\partial_\theta U_\theta]$.

Computation of $\partial_\theta \log Z_\theta$:

$$= \frac{\partial_\theta Z_\theta}{Z_\theta} = \int -\partial_\theta U_\theta(x) e^{-U_\theta(x)} \frac{1}{Z_\theta} dx = -\mathbb{E}_\theta[\partial_\theta U_\theta]$$

Gradient flow

$$\dot{\theta}(t) = \mathbb{E}_{\theta(t)}[\partial_\theta U_{\theta(t)}] - \mathbb{E}_*[\partial_\theta U_{\theta(t)}].$$

$\mathbb{E}_*[\partial_\theta U_\theta]$: is computed on the training samples $\approx \frac{1}{n} \sum_i \partial_\theta U_\theta(x_*^i)$
$\mathbb{E}_{\theta t}[\partial_\theta U_\theta]$: needs samples from the current model ρ_{θ_t}

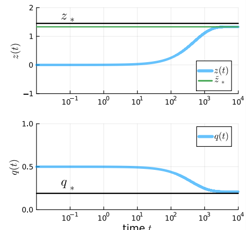
Proof of convergence.
$\partial_z U_z(x) = e^{-z}e^{-(x-b)^2/2}/U_z(x) \approx 1_{x \text{ is close to } b}$ hence

$$\forall z, w \qquad \mathbb{E}_w[\partial_z U_z] \approx \mathbb{P}_w(\text{ mode } b) = \frac{e^{-w}}{1+e^{-w}}$$

$$\dot{z}(t) \approx \frac{e^{-z(t)}}{1+e^{-z(t)}} - \frac{e^{-z_*}}{1+e^{-z_*}}.$$

Clearly this system converges towards its unique FP $z(t) = z_*$. $\square$

When estimating $\mathbb{E}_*$ using the samples x_*^i there can be a small correction: the empirical mass of mode b is replaced with $\hat{q}_* = \frac{e^{-\hat{z}_*}}{1+e^{-\hat{z}_*}}$ with $|\hat{z}_* - \hat{z}| = O(n^{-1/2})$.



MCMC sampling is too costly

Q: at each gradient step, how do we estimate $\mathbb{E}_\theta[\partial_\theta U_\theta]$?

A: using MCMC methods...

At step t, initialize X_0^i ("walkers"), then for $\tau = 0, \ldots, T_{mix}$,

$$X_{\tau+1}^i = X_\tau^i - \eta \nabla U_\theta(X_\tau^i) + \sqrt{2\eta}\xi_\tau$$

and estimate

$$\mathbb{E}_{\theta(t)}[\partial_\theta U_{\theta(t)}] \approx \frac{1}{N_{walkers}} \sum_{i=1}^{N_{walkers}} \partial_\theta U_{\theta(t)}(X_{T_{mix}}^i).$$

—

If T_{mix} is large, this is too costly. Each gradient ascent step will consume T_{mix} MCMC sampling steps for each of the $N_{walkers}$ chains!

Contrastive Divergence with k steps (CD-k), Hinton 2005

- don't let the chain reach T_{mix} steps. Use only k steps ($k = 1$).
- initialize each chain directly at the training points $\{x_*^i\}$.

Let $\tilde{\mathbb{P}}_\theta$ be the distribution of the negative samples. The Gradient Flow becomes

$$\dot{\theta}(t) = \tilde{E}_{\theta(t)}[\partial_\theta U_{\theta(t)}] - \mathbb{E}_*[\partial_\theta U_{\theta(t)}].$$

[Hyvarinen 2007]
in the limit of small noise $\eta \to 0$, CD-1 = score matching.

[Yair and Michaeli 20] CD-1 is an adversarial game

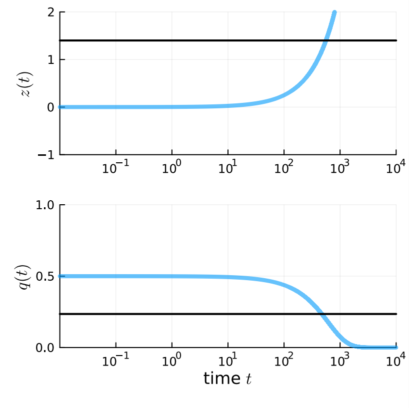
Persistent Contrastive Divergence (PCD), [Tieleman 2008]

- don't let the chain reach T_{mix} steps. Use only k steps ($k = 1$).
- ~~Initialize each chain directly at the training points $\{x_*^i\}$.~~
- initialize each chain directly where the previous chain ended.

Practically: maintain a set of *walkers* X_t^i. At step $t+1$,
1) approximate $\mathbb{E}_{\theta_t}[\partial_\theta U_{\theta(t)}] \approx \frac{1}{n}\sum_{i=1}^{N} \partial_\theta U_{\theta(t)}(X_t^i)$,
2) compute θ_{t+1} using the approximation,
3) move the walkers with $X_{t+1} = X_t - \eta \nabla U_{\theta(t+1)}(X_t) + \sqrt{2\eta}\xi$

Let $\hat{\mathbb{P}}_{\theta(t)}$ be the distribution of X_t. The gradient flow becomes

$$\dot{\theta}(t) = \hat{\mathbb{E}}_{\theta(t)}[\partial_\theta U_{\theta(t)}] - \mathbb{E}_*[\partial_\theta U_{\theta(t)}].$$



Mode collapse: one of the two modes disappears

Proof of mode collapse.
In theory the X_t^i should match the model $\rho_{z(t)}$. This is false in general!

$$\nabla U_z(x) \approx (x-a)1_{x \text{ close to } a} + (x-b)1_{x \text{ close to } b}$$

so if X_t is close to b, $dX_t \approx -(X_t - b)dt + \sqrt{2}dB_t$: this is an Ornstein-Uhlenbeck process centered at b. The two modes are stable.

There is no transfer of walkers from one mode to the other.

The distribution of X_t does not change and is equal to $\rho_{z(0)}$, hence te system becomes

$$\dot{z}(t) \approx \frac{e^{-z(0)}}{1+e^{-z(0)}} - \frac{e^{-z_*}}{1+e^{-z_*}}.$$

This leads to mode collapse, $z(t) \to \pm\infty$. $\square$

Reweighting PCD with Jarzynski's identity

Searching for the reweighting

Let U_t be any family of evolving potentials (such as U_{θ_t} given above). Consider the dynamics

$$dX_t = -\nabla U_t(X_t)dt + \sqrt{2}dB_t$$

Note $\hat{\rho}_t$ the law of X_t and $\rho_t = e^{-U_t}/Z_t$.

$$\partial_t \hat{\rho}_t = \Delta \hat{\rho}_t - \nabla \cdot (\nabla U_t \hat{\rho}_t)$$

ρ_t also solves this Fokker-Planck equation, hence $\rho_t = \hat{\rho}_t$ only at equilibrium; in general $\rho_t \neq \hat{\rho}_t$.

What is $\frac{d\rho_t}{d\hat{\rho}_t}$?

Jarzynski's augmented system

We add an auxiliary weight W_t to the system:

$$dX_t = -\nabla U_t(X_t)dt + \sqrt{2}dB_t \qquad X_0 \sim \rho_0 \tag{1}$$

$$dW_t = -W_t \dot{U}_t(X_t)dt \qquad W_0 = 1 \tag{2}$$

Note that W_t is an explicit path integral: $W_t = \exp\left\{-\int_0^t \dot{U}_s(X_s)ds\right\}$.

Theorem (Jarzynski reweighting)

$$\frac{\mathbb{E}[\varphi(X_t)W_t]}{\mathbb{E}[W_t]} = \mathbb{E}_{Y_t \sim \rho_t}[\varphi(Y_t)]$$

First appearance: for the computation of Z_t/Z_0, [Jarzynski 1996]

Proof outline

$\rho_t(x, w)$ = density of (X_t, W_t)

Define $\mu_t(x) = \int_0^\infty w\rho_t(x, w)dxdw$, so that

$$\mathbb{E}[\varphi(X_t)W_t] = \int \varphi(x)\mu_t(x)dx$$

1. Use Fokker-Planck for (4)-(5) to get

$$\dot{\mu}_t = \nabla \cdot (\nabla U_t \mu_t + \nabla \mu_t) + \dot{U}_t \mu_t \tag{3}$$

2. Check that $\rho_t = e^{-U_t - \log Z_t}$ also solves (3)
3. Unicity of solutions of parabolic PDEs

Discrete version

With a discrete family of evolving potentials U_k, define

$$\alpha_k(x,y) = U_k(x) + \frac{1}{2}(y-x)\cdot\nabla U_k(x) + \frac{1}{4}|\nabla U_k(x)|^2$$

The augmented system is:

$$X_{k+1} - X_k = -\varepsilon\nabla U_k(X_k) + N(0,\sqrt{2\varepsilon}) \qquad X_0 \sim \rho_0 \qquad (4)$$

$$W_{k+1} = W_k e^{\alpha_{k+1}(X_{k+1},X_k)+\alpha_k(X_k,X_{k+1})} \qquad W_0 = 1 \qquad (5)$$

Same result holds:

$$\frac{\mathbb{E}[\varphi(X_k)W_k]}{\mathbb{E}[W_k]} = \mathbb{E}_{Y_k\sim\rho_k}[\varphi(Y_k)]$$

Algorithm 1 Sequential Monte-Carlo training with Jarzynski correction

$A_0^i = 1$ for $i = 1, \ldots, N$.
for $k = 0, \ldots, K-1$ **do**
    $\bar{W}_k^i = W_k^i / \sum_{j=1}^N W_k^i$
    $\nabla_k = \sum_{i=1}^N \bar{W}_k^i \partial_\theta U_{\theta_k}(X_k^i) - n^{-1} \sum_{j=1}^n \partial_\theta U_{\theta_k}(x_*^j)$ ▷ gradient
    $\theta_{k+1} = \mathrm{opt}(\theta_k, \nabla_k)$ ▷ optimizer
    for $i = 1, \ldots, N$ **do**
        $X_{k+1}^i = X_k^i - h \nabla U_{\theta_k}(X_k^i) + \sqrt{2h}\, \xi_k^i$ ▷ ULA
        $W_{k+1}^i = W_k^i e^{\alpha_{k+1}(X_{k+1}^i, X_k^i) + \alpha_k(X_k^i, X_{k+1}^i)}$ ▷ update weight
    Resampling step (optional).

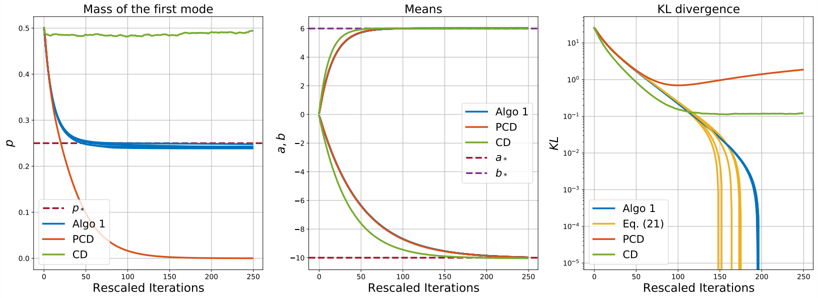


Figure 1: Learning also the modes

Some references

How to train your EBMs (Song & Kingma)

Improved CD (Du et al.)

Reduce, Reuse, Recycle (Du et al.)

Annealed Importance Sampling (Neal)