Energy-based generator matching: A neural sampler for general state space

Dongyeop Woo[1], Minsu Kim[1,2], Minky Kim[1], Kiyoung Seong[1], Sungsoo Ahn[1]
[1]KAIST, Korea Advanced Institute of Science and Technology , [2]Mila – Quebec AI Institute

GitHub PDF

NEURAL INFORMATION PROCESSING SYSTEMS

KAIST Mila

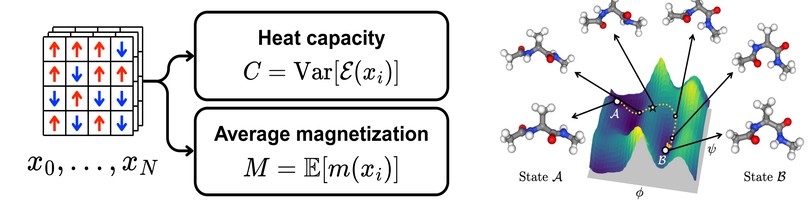
TL;DR Energy-driven training of neural samplers of **any Markov process** (diffusion/flow/jump) on **any state space**.

Motivation

Goal: Learn a generative model p_θ directly from an unnormalized density, without data:

$$p_\theta(x) \approx \tilde{p}(x) \propto \exp(-\mathcal{E}(x)).$$

Applications: Physical systems' property prediction.



Limitation of prior works: mostly restricted to continuous spaces and diffusion models.

Contribution: We propose EGM, a **modality-agnostic energy-driven training** scheme for **arbitrary Markov processes**.

Preliminary: generator matching

Generator is an operator describing the infinitesimal evolution of diffusion, flow, or jump processes:

$$\mathcal{L}_t f(x) := \frac{d}{dh}|_{h=0} \mathbb{E}_{X_{t+h}\sim p_{t+h|t}(\cdot|x)}[f(X_{t+h})] = \langle \mathcal{K} f(x), F_t(x)\rangle.$$

In practice, it can be parameterized as a function like score, flow vector field, or transition matrix. We denote this function as F_t.

Generator matching: Learn the generator F_t that produces prescribed probability path p_t by minimizing the matching loss,

$$L_{\mathrm{GM}}(\theta) = \mathbb{E}_{x_1\sim\tilde{p},\, t\sim\mathrm{Unif},\, x_t\sim p_{t|1}}\left[\left\|F_t^\theta(x_t) - F_t(x_t)\right\|_2^2\right].$$

Noise injecting prior — Flow vector field or score

Marginalization trick: Let $F_{t|1}$ be a conditional generator for $p_{t|1}$. Then the marginal generator is $F_t(x_t) = \mathbb{E}_{x_1\sim p_{1|t}(\cdot|x_t)}[F_{t|1}(x_t|x_1)]$.

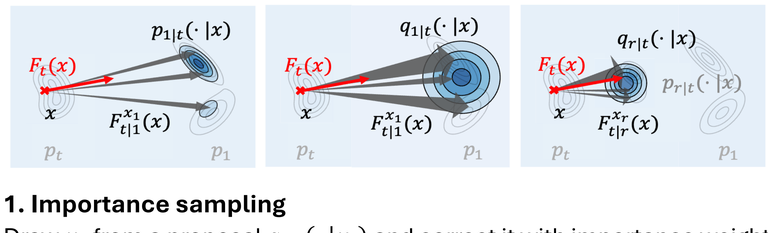
Method

We **estimate the marginal generator** using an unnormalized density and train the sampler with the the matching loss:

$$L_{\mathrm{EGM}}(\theta) = \mathbb{E}_{x_1\sim r,\, t\sim\mathrm{Unif},\, x_t\sim p_{t|1}}\left[\left\|F_t^\theta(x_t) - \widehat{F}_t(x_t)\right\|_2^2\right].$$

Collection policy r — Generator estimation

Overview of estimation schemes



1. Importance sampling

Draw x_1 from a proposal $q_{1|t}(\cdot|x_t)$ and correct it with importance weights:

$$\hat{F}_t(x_t) = \textstyle\sum_{x_1} w(x_1) F_{t|1}(x_t|x_1),$$

where $w(x_1) \propto p_{1|t}(x_1|x_t)/q_{1|t}(x_1|x_t)$ and $\sum w = 1$.

But the naïve $q_{1|t} \propto p_{t|1}$ has little overlap with $p_{1|t}$ → **incorrect estimation**.

2. Bootstrapping

Observation: When the time gap is small, $p_{t|r} \approx p_{r|t}$.

Estimation: Draw x_r from $q_{r|t} \propto p_{t|r}$, then estimate

$$\widehat{F}_t(x_t) = \textstyle\sum_{x_r} w(x_r) F_{t|r}(x_t|x_r),$$

where $w(x_r) \propto \tilde{p}_r(x_r)$ and $\sum w = 1$.

Surrogate energy: $\tilde{p}_r(x_r)$ is unknown → learn a surrogate energy model $\tilde{p}_r^\theta(x_r)$ using 'noised energy matching'.

Bootstrapping reduces the overall variance of the estimation!

Collection policy: iterative training with replay buffer

Outer loop: update buffer with model samples.

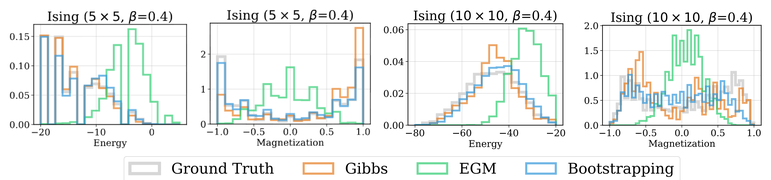
Inner loop: fit L_{EGM} with buffer samples.

Experiments

Discrete task

	5 x 5 Ising (d = 25)		10 x 10 Ising (d= 100)	
Algorithm ↓	$\mathcal{E}$-$\mathcal{W}_1$ ↓	M-$\mathcal{W}_1$ ↓	$\mathcal{E}$-$\mathcal{W}_1$ ↓	M-$\mathcal{W}_1$ ↓
Gibbs	0.71 ± 0.43	0.23 ± 0.18	5.29 ± 1.95	0.29 ± 0.11
EGM	3.73 ± 0.39	0.24 ± 0.02	19.94 ± 0.69	0.36 ± 0.01
+ Bootstrap	$\mathbf{0.60 \pm 0.12}$	$\mathbf{0.04 \pm 0.01}$	$\mathbf{2.51 \pm 0.16}$	$\mathbf{0.24 \pm 0.01}$

Energy & magnetization histogram



Multimodal task

	GB-RBM (d = 5)		JointMoG (d = 20)
Method ↓	$\mathcal{E}$-$\mathcal{W}_1$ ↓	x-$\mathcal{W}_2$ ↓	$\mathcal{E}$-$\mathcal{W}_1$ ↓
Gibbs	0.77 ± 0.03	5.12 ± 0.21	7.76 ± 0.02
EGM	0.33 ± 0.02	0.63 ± 0.10	8.83 ± 0.12
+Bootstrap	$\mathbf{0.20 \pm 0.15}$	$\mathbf{0.61 \pm 0.09}$	$\mathbf{1.20 \pm 0.32}$

Sample projection plot (left: GB-RBM, right: JointMoG)