Training Energy-Based Models with Energy Discrepancies

Zijing Ou
Imperial College London

18/10/2023

Hello

- Second year PhD at Imperial College London
- I work on a wide variety of topics in ML/Probabilistic Inference:
  - ❑ Energy-based modelling
  - ❑ Explainability
  - ❑ Representation Learning
  - ❑ Generative Models
  - ❑ …
- Today, I gonna talk about my recent research on training EBMs

Energy-based Models

energy function

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

normalising constant /
partition function

$$Z(\theta) = \int \exp[-E(x;\theta)]dx$$

Energy-based Models

energy function

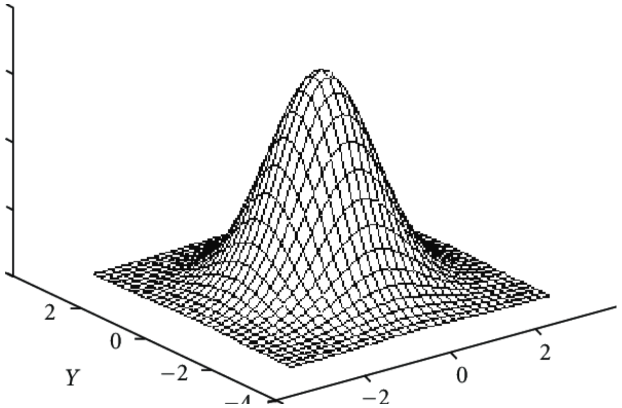
$$p_\theta(x) = \frac{1}{Z(\theta)} \exp[-E(x;\theta)]$$

normalising constant / partition function

$$Z(\theta) = \int \exp[-E(x;\theta)]dx$$

Examples: Gaussian (continuous)

- $E(x;\theta) = \frac{1}{2\sigma^2}(x-\mu)^2$
- $\theta = \{\mu, \sigma^2\}$
- $Z(\theta) = \sqrt{2\pi\sigma^2}$
- $x \in \mathbb{R}^{D_x}$



Energy-based Models

energy function

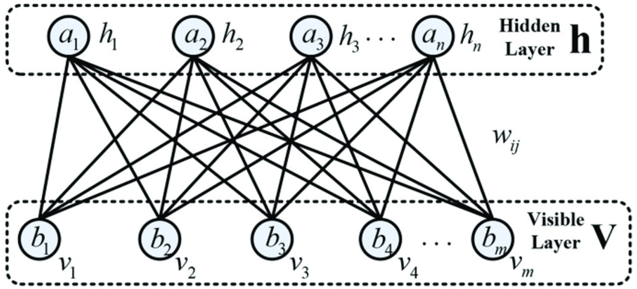
$$p_\theta(x) = \frac{1}{Z(\theta)} \exp[-E(x;\theta)]$$

normalising constant / partition function

$$Z(\theta) = \int \exp[-E(x;\theta)]dx$$

Examples: Restricted Boltzmann Machine (discrete)

- $-E(x;\theta) = b_x^T x + b_h^T h + x^T W h$
- $\theta = \{b_x^T, b_h^T, W\}$
- $Z = \sum_{x,h} \exp[b_x^T x + b_h^T h + x^T W h]$
- $x \in \{0,1\}^{D_x}, h \in \{0,1\}^{D_h}$



Training EBMs

Maximum Likelihood Estimation of θ:

$$\theta^* = \arg\max_{\theta} \mathbb{E}_{p_{data}(x)}[-E(x;\theta) - \log Z(\theta)]$$

$$-\nabla_\theta E_{p_{data}(x)}[\log p_\theta(x)] = \mathbb{E}_{p_{data}(x)}[\nabla_\theta E(x;\theta)] - \mathbb{E}_{p_\theta(x)}[\nabla_\theta E(x;\theta)]$$

decrease energy around data

increase energy around samples

Training EBMs

Maximum Likelihood Estimation of θ:

$$\theta^* = \arg\max_\theta \mathbb{E}_{p_{data}(x)}[-E(x;\theta) - \log Z(\theta)]$$

$$-\nabla_\theta E_{p_{data}(x)}[\log p_\theta(x)] = \mathbb{E}_{p_{data}(x)}[\nabla_\theta E(x;\theta)] - \mathbb{E}_{p_\theta(x)}[\nabla_\theta E(x;\theta)]$$

decrease energy around data — increase energy around samples

Examples: Restricted Boltzmann Machine

- $-E(x;\theta) = b_x^T x + b_h^T h + x^T W h$
- $-\nabla_\theta E_{p_{data}(x)}[\log p_\theta(x)] =$

$$E_{p_{data}(x)p_\theta(h|x)}[\nabla_\theta E(x,h;\theta)] - E_{p_\theta(x,h)}[\nabla_\theta E(x,h;\theta)]$$

sample h conditioned on data — simulate $h, x \sim p_\theta(x,h)$

Training EBMs

Maximum Likelihood Estimation of θ:

$$-\nabla_\theta \mathbb{E}_{p_{data}(x)}[\log p_\theta(x)] = \mathbb{E}_{p_{data}(x)}[\nabla_\theta E(x;\theta)] - \mathbb{E}_{p_\theta(x)}[\nabla_\theta E(x;\theta)]$$

Simulate $x \sim p_\theta(x)$ with Langevin dynamics

$$x_{t+1} = x_t - \eta \nabla_x E(x;\theta) + \sqrt{2\eta}\epsilon, \qquad \epsilon \sim N(0, I)$$

$$\eta \to 0, x_{t\to\infty} \sim p_\theta(x)$$

Hinton, Geoffrey E. Training products of experts by minimizing contrastive divergence. Neural Computing, 2002.

Training EBMs

Maximum Likelihood Estimation of θ:

$$-\nabla_\theta \mathbb{E}_{p_{data}(x)}[\log p_\theta(x)] = \mathbb{E}_{p_{data}(x)}[\nabla_\theta E(x;\theta)] - \mathbb{E}_{p_\theta(x)}[\nabla_\theta E(x;\theta)]$$

Simulate $x \sim p_\theta(x)$ with Langevin dynamics

$$x_{t+1} = x_t - \eta \nabla_x E(x;\theta) + \sqrt{2\eta}\epsilon, \qquad \epsilon \sim N(0, I)$$

$$\eta \to 0, x_{t\to\infty} \sim p_\theta(x)$$

Simulating MCMC is time-consuming!!!

Hinton, Geoffrey E. Training products of experts by minimizing contrastive divergence. Neural Computing, 2002.

Training EBMs

Minimising Fisher Divergence:

$$FD(p_{data}, p_\theta) = \mathbb{E}_{p_{data}(x)}[||\nabla_x \log p_{data}(x) - \nabla_x \log p_\theta(x)||^2]$$

Intractable term

This leads to the score-matching loss:

$$SM(p_{data}, p_\theta) = \mathbb{E}_{p_{data}(x)}\left[\frac{1}{2}||\nabla_x E_\theta(x)||^2 - Tr(\nabla_x^2 E_\theta(x))\right]$$

Hessian matrix

Hyvärinen et al. Estimation of non-normalized statistical models by score matching. JMLR, 2005.

Training EBMs

Minimising Fisher Divergence:

$$FD(p_{data}, p_\theta) = \mathbb{E}_{p_{data}(x)}[||\nabla_x \log p_{data}(x) - \nabla_x \log p_\theta(x)||^2]$$

Intractable term

This leads to the score-matching loss:

$$SM(p_{data}, p_\theta) = \mathbb{E}_{p_{data}(x)}\left[\frac{1}{2}||\nabla_x E_\theta(x)||^2 - Tr(\nabla_x^2 E_\theta(x))\right]$$

Hessian matrix

MCMC-free, BUT require Second-Order Computation

Hyvärinen et al. Estimation of non-normalized statistical models by score matching. JMLR, 2005.

In This Work

We propose **Energy Discrepancy**, a score-independent loss for training EBMs

Given the **contrastive potential** induced by conditional density $q(y|x)$ as

$$U_q(y) := -\log \sum_{x \in \mathcal{X}} q(y|x) \exp(-U(x))$$

We define the **energy discrepancy** between p_{data} and U induced by q as

$$ED_q(p_{data}, U) = \mathbb{E}_{p_{data}(x)}[U(x)] - \mathbb{E}_{p_{data}(x)} \mathbb{E}_{q(y|x)}[U_q(y)]$$

Tobias, Ou, et al. Energy Discrepancies: A Score-Independent Loss for Energy-Based Models. NeurIPS, 2023.

In This Work

We propose **Energy Discrepancy**, a score-independent loss for training EBMs

Given the **contrastive potential** induced by conditional density $q(y|x)$ as

$$U_q(y) := -\log \sum_{x \in \mathcal{X}} q(y|x) \exp(-U(x))$$

We define the **energy discrepancy** between p_{data} and U induced by q as

$$ED_q(p_{data}, U) = \mathbb{E}_{p_{data}(x)}[U(x)] - \mathbb{E}_{p_{data}(x)} \mathbb{E}_{q(y|x)}[U_q(y)]$$

Non-Parametric Estimation Results

$$U^* := \operatorname*{argmin}_U ED_q(p_{data}, U) \quad \Rightarrow \quad p_{data}(x) \propto \exp(-U^*(x))$$

Tobias, Ou, et al. Energy Discrepancies: A Score-Independent Loss for Energy-Based Models. NeurIPS, 2023.

Connections to Other Methods

Connection to the **KL-Contraction Divergence**

Denote the convolution operator as

$$\mathcal{Q}p(y) \coloneqq \sum_{x\in\mathcal{X}} q(y|x)p(x)$$

The KL-Contraction Divergence constructs a valid objective

$$KLC_{\mathcal{Q}}(p_1, p_2) = KL(p_1 \| p_2) - KL(\mathcal{Q}p_1 \| \mathcal{Q}p_2)$$

$KLC_{\mathcal{Q}}(p_1, p_2) \geq 0$ and $= 0$ iff $p_1 = p_2, a.e.$

Connections to Other Methods

Connection to the **KL-Contraction Divergence**

Denote the convolution operator as

$$Qp(y) \coloneqq \sum_{x \in \mathcal{X}} q(y|x)p(x)$$

The KL-Contraction Divergence constructs a valid objective

$$KLC_Q(p_1, p_2) = KL(p_1 \| p_2) - KL(Qp_1 \| Qp_2)$$

$KLC_Q(p_1, p_2) \geq 0$ and $= 0$ iff $p_1 = p_2, a.e.$

Connections to Energy Discrepancy

$$\underset{U}{\operatorname{argmin}}\, ED_q(p_{data}, U) \quad \Leftrightarrow \quad \underset{U}{\operatorname{argmin}}\, KLC_Q(p_{data}, p_{ebm}),\ p_{ebm} \propto \exp(-U)$$

Connections to Other Methods

Connection to the **Maximum Recovery Likelihood**

Denote the posterior $p_{ebm}(x|y)$ as

$$p_{ebm}(x|y) \propto \exp\big(-U(x)\big)q(y|x)$$

The Maximum Recovery Likelihood constructs a valid objective

$$RL_q(p_{data}, U) = \mathbb{E}_{p_{data}(x)}\mathbb{E}_{q(y|x)}[\log p_{ebm}(x|y)]$$

Tobias, Ou, et al. Energy Discrepancies: A Score-Independent Loss for Energy-Based Models. NeurIPS, 2023.

Connections to Other Methods

Connection to the **Maximum Recovery Likelihood**

Denote the posterior $p_{ebm}(x|y)$ as

$$p_{ebm}(x|y) \propto \exp\big(-U(x)\big)q(y|x)$$

The Maximum Recovery Likelihood constructs a valid objective

$$RL_q(p_{data}, U) = \mathbb{E}_{p_{data}(x)}\mathbb{E}_{q(y|x)}[\log p_{ebm}(x|y)]$$

Connections to Energy Discrepancy

$$\operatorname*{argmin}_U ED_q(p_{data}, U) \quad \Leftrightarrow \quad \operatorname*{argmin}_U -RL_q(p_{data}, U)$$

Tobias, Ou, et al. Energy Discrepancies: A Score-Independent Loss for Energy-Based Models. NeurIPS, 2023.

Connections to Other Methods

Connection to the **Contrastive Divergence**

Assume the conditional density $q(y|x)$ satisfies the detailed balance

$$q(y|x)\exp\big(-U(x)\big) = q(x|y)\exp\big(-U(y)\big)$$

The Contrastive Divergence constructs a valid objective

$$CD(p_{data}, U) = \mathbb{E}_{p_{data}(x)}[U(x)] - \mathbb{E}_{p_{ebm}(x)}[U(x)]$$

Tobias, Ou, et al. Energy Discrepancies: A Score-Independent Loss for Energy-Based Models. NeurIPS, 2023.

Connections to Other Methods

Connection to the **Contrastive Divergence**

Assume the conditional density $q(y|x)$ satisfies the detailed balance

$$q(y|x)\exp\big(-U(x)\big) = q(x|y)\exp\big(-U(y)\big)$$

The Contrastive Divergence constructs a valid objective

$$CD(p_{data}, U) = \mathbb{E}_{p_{data}(x)}[U(x)] - \mathbb{E}_{p_{ebm}(x)}[U(x)]$$

Connections to Energy Discrepancy

$$\operatorname*{argmin}_U ED_q(p_{data}, U) \quad \Leftrightarrow \quad \operatorname*{argmin}_U -CD(p_{data}, U)$$

Tobias, Ou, et al. Energy Discrepancies: A Score-Independent Loss for Energy-Based Models. NeurIPS, 2023.

Energy Discrepancy In Practice

$$ED_q(p_{data}, U) = \mathbb{E}_{p_{data}(x)}[U(x)] - \mathbb{E}_{p_{data}(x)}\mathbb{E}_{q(y|x)}[U_q(y)]$$

with the contrastive potential defined as

$$U_q(y) = -\log \sum_{x \in \mathcal{X}} q(y|x) \exp(-U(x))$$

Energy Discrepancy In Practice

$$ED_q(p_{data}, U) = \mathbb{E}_{p_{data}(x)}[U(x)] - \mathbb{E}_{p_{data}(x)}\mathbb{E}_{q(y|x)}[U_q(y)]$$

with the contrastive potential defined as

$$U_q(y) = -\log \sum_{x \in \mathcal{X}} q(y|x) \exp(-U(x))$$

Estimating U_q with Importance Sampling

$$U_q(y) = -\mathbb{E}_{\rho_y(x)}\left[\frac{q(y|x)}{\rho_y(x)} \exp(-U(x))\right]$$

Energy Discrepancy In Practice

$$ED_q(p_{data}, U) = \mathbb{E}_{p_{data}(x)}[U(x)] - \mathbb{E}_{p_{data}(x)}\mathbb{E}_{q(y|x)}[U_q(y)]$$

with the contrastive potential defined as

$$U_q(y) = -\log \sum_{x \in \mathcal{X}} q(y|x) \exp(-U(x))$$

Estimating U_q with Importance Sampling

$$U_q(y) = -\mathbb{E}_{\rho_y(x)}\left[\frac{q(y|x)}{\rho_y(x)} \exp(-U(x))\right]$$

A simple choice of $\rho_y(x)$ is an uninformed proposal

$$\rho_y(x) := \frac{q(y|x)}{\sum_{x \in \mathcal{X}} q(y|x)}$$

$\rho_y(x)$ is tractable for some perturbations, e.g., Gaussian, Bernoulli, etc.

Continuous Energy Discrepancy

Let q_t be the density involved by the diffusion process

$$dx_t = a(x_t)dt + dw_t$$

The energy discrepancy is given by a **multi-noise score matching** loss

$$ED_{q_t}(p_{data}, U) = \int_0^t \underbrace{\mathbb{E}_{p_s(x_s)}\left[\frac{1}{2}\left\|\nabla_{x_s} U_{q_s}(x_s)\right\|^2 - Tr(\nabla_x^2 U_{q_s}(x_s))\right]}_{:= SM(p_s, U_{q_s})} ds + const$$

$$p_s(y) := \int q_s(y|x) p_{data}(x) dx, \;\; \exp\left(-U_{q_s}(y)\right) := \int q_s(y|x) \exp(-U(x)) dx$$

Tobias, Ou, et al. Energy Discrepancies: A Score-Independent Loss for Energy-Based Models. NeurIPS, 2023.

Continuous Energy Discrepancy

Let q_t be the density involved by the diffusion process

$$dx_t = a(x_t)dt + dw_t$$

The energy discrepancy is given by a **multi-noise score matching** loss

$$ED_{q_t}(p_{data}, U) = \int_0^t \underbrace{\mathbb{E}_{p_s(x_s)}\left[\frac{1}{2}\left\|\nabla_{x_s} U_{q_s}(x_s)\right\|^2 - Tr\left(\nabla_x^2 U_{q_s}(x_s)\right)\right]}_{:= SM(p_s, U_{q_s})} ds + const$$

$$p_s(y) := \int q_s(y|x) p_{data}(x) dx, \quad \exp\left(-U_{q_s}(y)\right) := \int q_s(y|x) \exp\left(-U(x)\right) dx$$

If $a = 0$, energy discrepancy converges to **maximum likelihood** if $t \to +\infty$

$$\left| ED_{q_t}(p_{data}, U) + \mathbb{E}_{p_{data}(x)}[\log p_{ebm}(x)] - c(t) \right| \leq \frac{1}{2t} \mathbb{W}_2^2(p_{data}, p_{ebm})$$

$\mathbb{W}(\cdot,\cdot)$ denotes the Wasserstein distance

Tobias, Ou, et al. Energy Discrepancies: A Score-Independent Loss for Energy-Based Models. NeurIPS, 2023.

Continuous Energy Discrepancy

Connections to score matching and maximum likelihood

$$p_\rho(x) = \rho g_1(x) + (1-\rho) g_2(x)$$

Energy Discrepancy under different t

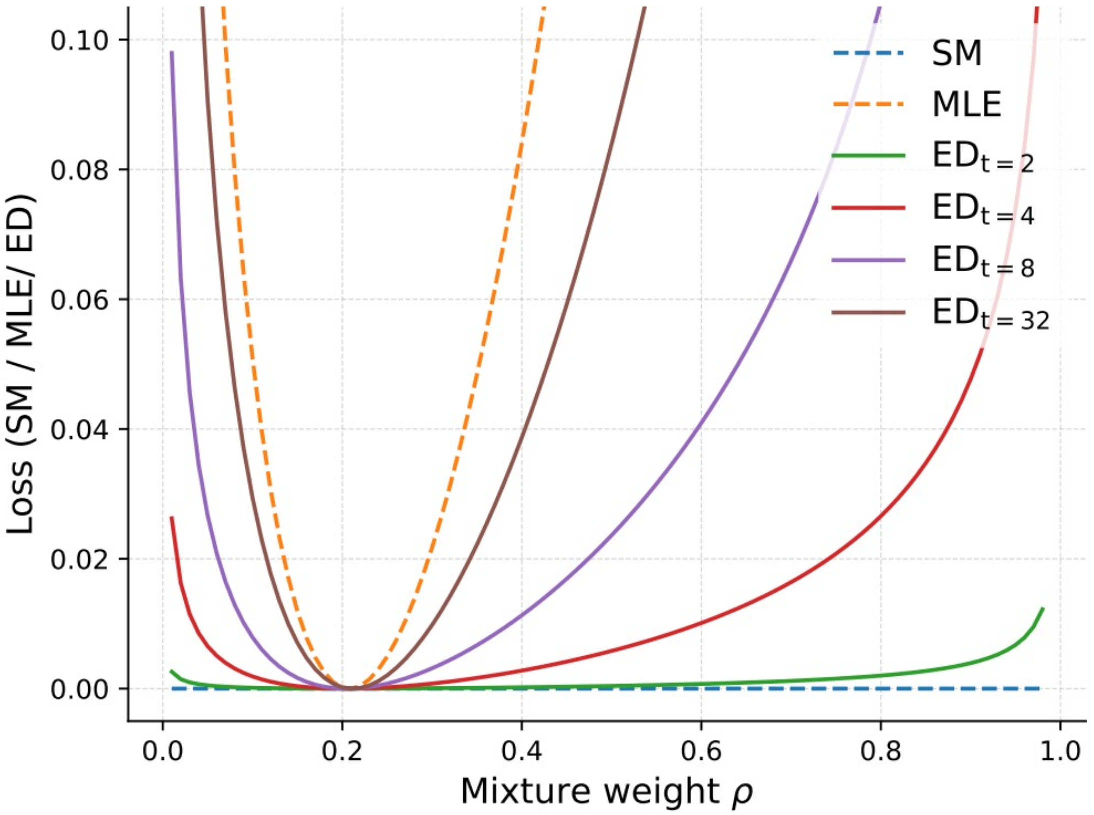
$$ED_{q_t}(p_{\rho=0.2}, \log p_\rho)$$

Score Matching

$$\frac{1}{2}\mathbb{E}_{p_{\rho=0.2}(x)}[||\nabla \log p_{\rho=0.2}(x) - \nabla \log p_\rho\,(x)||^2]$$

Maximum Likelihood

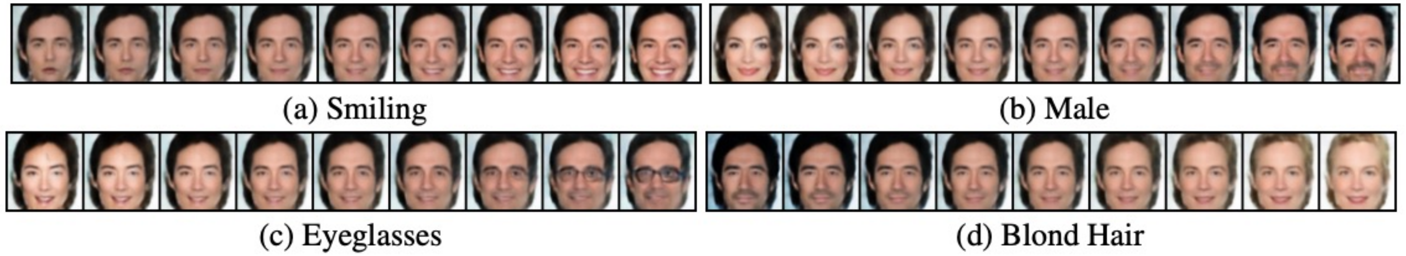
$$\mathbb{E}_{p_{\rho=0.2}(x)}[-\log p_\rho(x)]$$



Tobias, Ou, et al. Energy Discrepancies: A Score-Independent Loss for Energy-Based Models. NeurIPS, 2023.

Continuous Energy Discrepancy

Learning laten EBMs



Attribute Manipulate

$$p_{\phi,\theta}(x) \propto \int p_{\phi}(x|z)\exp\big(-\mathrm{E}_{\theta}(\mathrm{z})\big)\mathrm{dz}$$

$$p_{\phi,\theta}(z|x) \propto p_{\phi}(x|z)\exp\big(-\mathrm{E}_{\theta}(\mathrm{z})\big)$$

Unconditional Generation

Discrete Energy Discrepancy

$$U_q(y) \coloneqq -\log \sum_{x \in \mathcal{X}} q(y|x) \exp(-U(x))$$

$$ED_q(p_{data}, U) = \mathbb{E}_{p_{data}(x)}[U(x)] - \mathbb{E}_{p_{data}(x)} \mathbb{E}_{q(y|x)}[U_q(y)]$$

Energy discrepancy is valid in discrete spaces $\mathcal{X} \in \{0,1\}^d$

We can define $q(y|x)$ as Bernoulli perturbation

$$y = x + \xi \; mod \; 2, \xi \sim Bernoulli(\epsilon)^d, \; \epsilon \in (0,1)$$

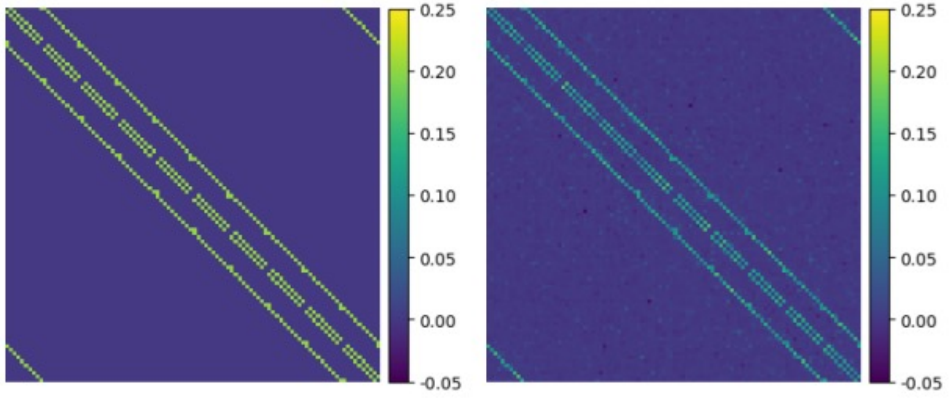
Discrete Energy Discrepancy

$$U_q(y) := -\log \sum_{x \in \mathcal{X}} q(y|x) \exp(-U(x))$$

$$ED_q(p_{data}, U) = \mathbb{E}_{p_{data}(x)}[U(x)] - \mathbb{E}_{p_{data}(x)} \mathbb{E}_{q(y|x)}[U_q(y)]$$

Energy discrepancy is valid in discrete spaces $\mathcal{X} \in \{0,1\}^d$

Applications: training Ising models



Ground Truth　　　Learned Pattern

Tobias, Ou, et al. Training Discrete EBMs with Energy Discrepancy. ICML Workshop SODS, 2023.

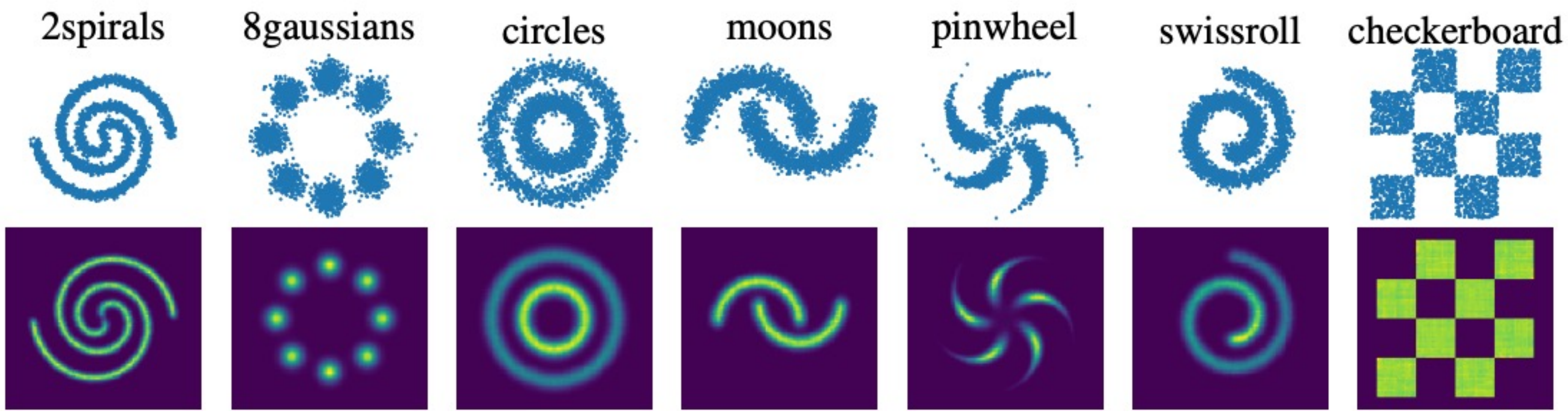
Discrete Energy Discrepancy

$$U_q(y) := -\log \sum_{x\in\mathcal{X}} q(y|x)\exp(-U(x))$$

$$ED_q(p_{data}, U) = \mathbb{E}_{p_{data}(x)}[U(x)] - \mathbb{E}_{p_{data}(x)}\mathbb{E}_{q(y|x)}[U_q(y)]$$

Energy discrepancy is valid in discrete spaces $\mathcal{X} \in \{0,1\}^d$

Applications: density estimation



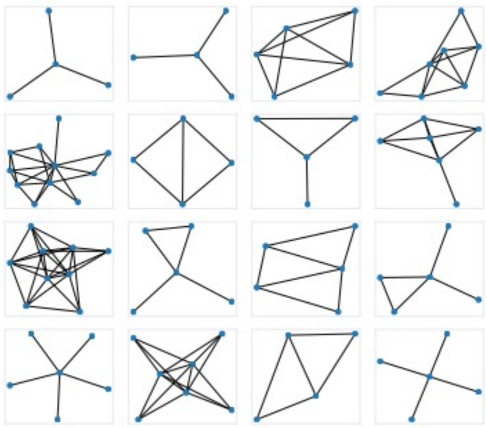
Discrete Energy Discrepancy

$$U_q(y) := -\log \sum_{x\in\mathcal{X}} q(y|x)\exp(-U(x))$$

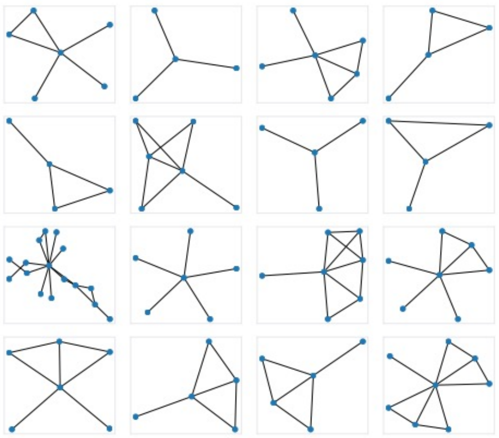
$$ED_q(p_{data}, U) = \mathbb{E}_{p_{data}(x)}[U(x)] - \mathbb{E}_{p_{data}(x)}\mathbb{E}_{q(y|x)}[U_q(y)]$$

Energy discrepancy is valid in discrete spaces $\mathcal{X} \in \{0,1\}^d$

Applications: ego-graph generation



Training Data



Generated Data

Tobias, Ou, et al. Training Discrete EBMs with Energy Discrepancy. ICML Workshop SODS, 2023.

Thank you!

Questions? Ask now, or email:
z.ou22@imperial.ac.uk

Thanks to my awesome collaborators:


Tobias Schröder


Jen Ning Lim


Yingzhen Li


Sebastian Vollmer


Andrew Duncan