The Modern Arts of Discrete Energy-based Models Training and Inference

Zijing Ou
CSML Reading Groups

03/30/2023

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

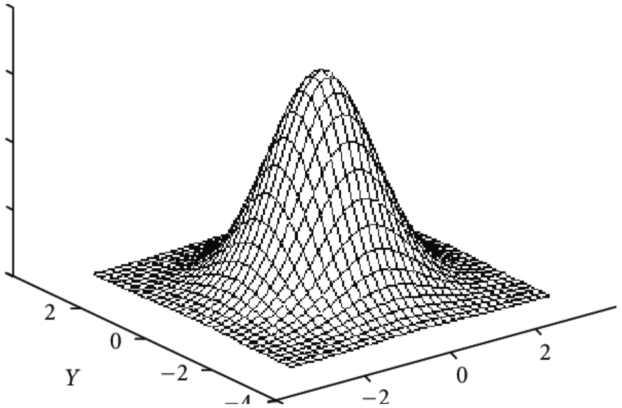
$$p_\theta(x) = \frac{1}{Z(\theta)} \exp[-E(x;\theta)]$$

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

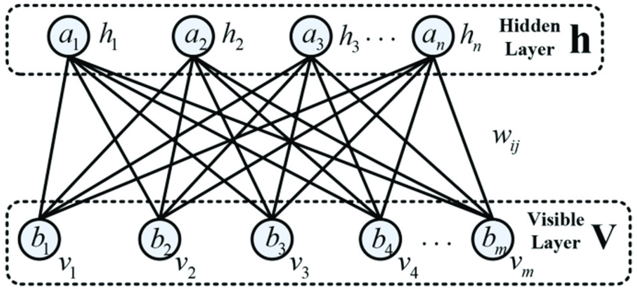
$$p_\theta(x) = \frac{1}{Z(\theta)} \exp[-E(x;\theta)]$$

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:

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

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

decrease energy around data

increase energy around samples

Training EBMs

Maximum Likelihood Estimation of θ:

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

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

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:

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

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

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

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

Training EBMs

Maximum Likelihood Estimation of θ:

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

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

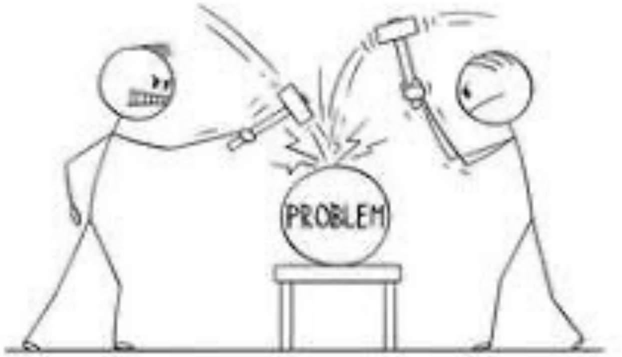
$$x_{t+1} = x_t - \eta \nabla_x E(x;\theta) + \sqrt{2\eta}\epsilon, \qquad \epsilon \sim N(0, I)$$

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

How to simulate samples on DISCRETE space? 🤔

The Family of Locally Balanced Samplers

Locally Balanced Samplers



Discrete EBMs Training

Locally Informed Proposals

Metropolis Hastings Sampler

proposal distribution

$$\min\left\{1, \exp\big(f(x') - f(x)\big)\frac{q(x|x')}{q(x'|x)}\right\}$$

Locally Informed Proposals

Metropolis Hastings Sampler

proposal distribution

$$\min\left\{1, \exp\big(f(x') - f(x)\big)\frac{q(x|x')}{q(x'|x)}\right\}$$

Locally-informed proposals

$$q(x'|x) \propto g\Big(\exp\big(f(x') - f(x)\big)\Big)K_\sigma(x'|x)$$

- $K(x'|x)$: a uniform distribution over a local ball of radius σ
- $\exp(f(x') - f(x))$: reweight the uninformed kernel according to the target disctribution
- $g(t) = \sqrt{t}$: balancing function balances the acceptance and rejection probabilities

Zanella Giacomo. Informed proposals for local MCMC in discrete spaces. Journal of the American Statistical Association, 2020.

Gibbs with Gradients

Locally-informed proposals

$$q(x'|x) \propto \exp\left(f(x') - f(x)\right) \mathbb{I}_{x' \in \mathcal{N}(x)}$$

Common discrete distributions are defined on the top of continuous distributions

Distribution	$\log p(x) + \log Z$
Categorical	$x^T\theta$
Poisson[1]	$x \log \lambda - \log \Gamma(x+1)$
HMM	$\sum_{t=1}^{T} x_{t+1}^T A x_t - \frac{(w^T x - y)^2}{2\sigma^2}$
RBM	$\sum_i \text{softplus}(Wx+b)_i + c^T x$
Ising	$x^T W x + b^T x$
Potts	$\sum_{i=1}^{L} h_i^T x_i + \sum_{i,j=1}^{L} x_i^T J_{ij} x_j$
Deep EBM	$f_\theta(x)$

Grathwohl, et al. Oops i took a gradient: Scalable sampling for discrete distributions. ICML, 2021.

Gibbs with Gradients

Locally-informed proposals

$$q(x'|x) \propto \exp\big(f(x') - f(x)\big)\, \mathbb{I}_{x' \in \mathcal{N}(x)}$$

Gibbs with Gradients

$$q(x'|x) \propto \exp\left(\frac{1}{2} \nabla_x f(x)^T (x' - x)\right) \mathbb{I}_{x' \in \mathcal{N}(x)}$$

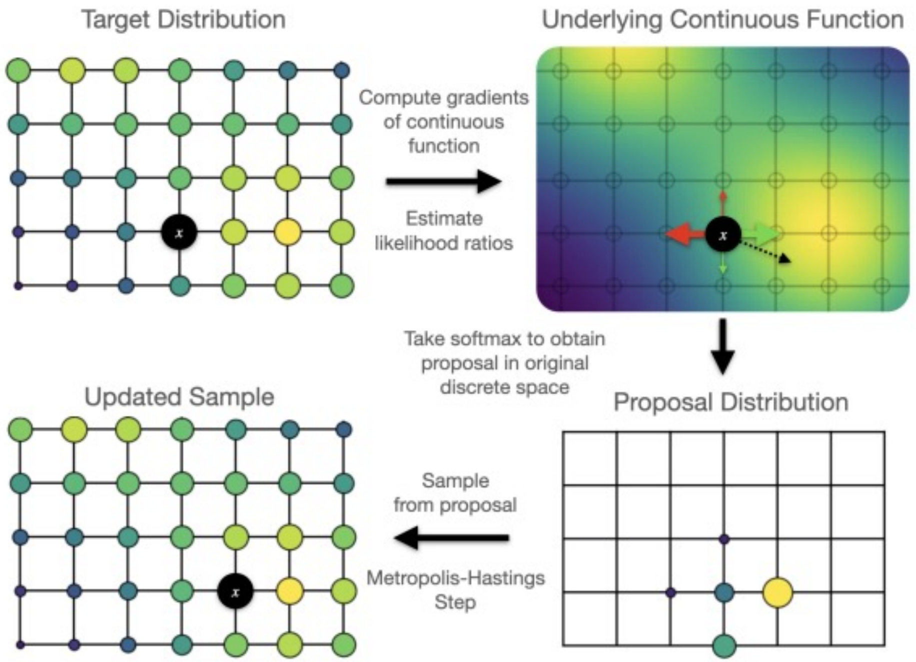
computational complexity:

$$\mathcal{O}(D) \to \mathcal{O}(1)$$

Gibbs with Gradients

Gibbs with Gradients

$$q(x'|x) \propto \exp\left(\frac{1}{2}\nabla_x f(x)^T (x' - x)\right) \mathbb{I}_{x' \in \mathcal{N}(x)}$$



Grathwohl, et al. Oops i took a gradient: Scalable sampling for discrete distributions. ICML, 2021.

Path Auxiliary Proposals

Gibbs with Gradients

1-Hamming ball

$$q(x'|x) \propto \exp\left(\frac{1}{2}\nabla_x f(x)^T (x' - x)\right) \mathbb{I}_{x' \in \mathcal{N}(x)}$$

GwG updates 1 bit per MH step => Could we update multi-bits per step?

Path Auxiliary Proposals

Gibbs with Gradients

1-Hamming ball

$$q(x'|x) \propto \exp\left(\frac{1}{2}\nabla_x f(x)^T(x'-x)\right)\mathbb{I}_{x'\in\mathcal{N}(x)}$$

GwG updates 1 bit per MH step => Could we update multi-bits per step?

Naïve Solution: increase the window-size of Hamming ball

1-Hamming ball → K-Hamming ball

$$\mathcal{O}(1) \to \mathcal{O}(D^K)$$

Path Auxiliary Proposals

Example: a auxiliary path of length 3

AAA → AAB → AAC → ABC

Path Auxiliary Proposals

$$
\begin{aligned}
q_K(x'|x) &= \prod_{k=1}^{K} q\left(x^k \middle| x^{k-1}\right) \\
&\propto \prod_{k=1}^{K} \exp\left(\frac{1}{2}\nabla_x f(x)^T (x^k - x^{k-1})\right) \mathbb{I}_{x^k \in \mathcal{N}(x^{k-1})}
\end{aligned}
$$

K -Hamming ball → K-length path

$\mathcal{O}(D^K) \rightarrow \mathcal{O}(K)$

Sun, et al. Path auxiliary proposal for mcmc in discrete space. ICLR, 2022.
Patrick, et al. Plug & Play Directed Evolution of Proteins with Gradient-based Discrete MCMC. Arxiv, 2022.

Path Auxiliary Proposals

Path Auxiliary Proposals

$$q_K(x'|x) = \prod_{k=1}^{K} q\left(x^k \middle| x^{k-1}\right)$$
$$\propto \prod_{k=1}^{K} \exp\left(\frac{1}{2}\nabla_x f(x)^T (x^k - x^{k-1})\right) \mathbb{I}_{x^k \in \mathcal{N}(x^{k-1})}$$

PAPs with 1 step from K-Hamming ball

vs GwGs with K steps from 1-Hamming ball

$$\min\left\{1, \frac{f(x^K) \prod_{k=1}^{K} q(x^k|x^{k-1})}{f(x) \prod_{k=1}^{K} q(x^{k-1}|x^k)}\right\} \geq \prod_{k=1}^{K} \min\left\{1, \frac{f(x^K) q(x^k|x^{k-1})}{f(x) q(x^{k-1}|x^k)}\right\}$$

Sun, et al. Path auxiliary proposal for mcmc in discrete space. ICLR, 2022.
Patrick, et al. Plug & Play Directed Evolution of Proteins with Gradient-based Discrete MCMC. Arxiv, 2022.

Path Auxiliary Proposals

Path Auxiliary Proposals

$$q_K(x'|x) = \prod_{k=1}^{K} q\left(x^k \middle| x^{k-1}\right)$$
$$\propto \prod_{k=1}^{K} \exp\left(\frac{1}{2}\nabla_x f(x)^T (x^k - x^{k-1})\right) \mathbb{I}_{x^k \in \mathcal{N}(x^{k-1})}$$

Optimal choice of K

Theorem: *The optimal choice of scale for* $K = lD_x^{2/3}$ *is obtained when the expected acceptance is* 0.574*, independent of the target distribution.*

The optimal acceptance rates for random walk Metropolis is 0.234*.*

Discrete Langevin Proposals

Locally-informed proposals

$$q(x'|x) \propto \exp\left(\frac{1}{2}f(x') - \frac{1}{2}f(x)\right) K_\sigma(x'|x)$$

PAP updates multi-bits per MH step => Could we update all-bits in parallel?

Discrete Langevin Proposals

Locally-informed proposals

$$q(x'|x) \propto \exp\left(\frac{1}{2}f(x') - \frac{1}{2}f(x)\right) K_\sigma(x'|x)$$

PAP updates multi-bits per MH step => Could we update all-bits in parallel?

$$q(x'|x) \propto \exp\left(\frac{1}{2}f(x') - \frac{1}{2}f(x)\right) \exp\left(-\frac{||x'-x||^2}{2\alpha}\right)$$

$$f(x') - f(x) \approx \nabla_x f(x)^T (x' - x)$$

Zhang, et al. A Langevin-like sampler for discrete distributions. ICML 2022.
Benjamin, et al. Enhanced gradient-based MCMC in discrete spaces. TMLR 2022.

Discrete Langevin Proposals

Discrete Langevin Proposals

$$q(x'|x) \propto \exp\left(-\frac{1}{2\alpha}\left\|x' - x - \frac{\alpha}{2}\nabla_x f(x)\right\|^2\right)$$

Zhang, et al. A Langevin-like sampler for discrete distributions. ICML 2022.
Benjamin, et al. Enhanced gradient-based MCMC in discrete spaces. TMLR 2022.

Discrete Langevin Proposals

Discrete Langevin Proposals

$$q(x'|x) \propto \exp\left(-\frac{1}{2\alpha}\left\|x' - x - \frac{\alpha}{2}\nabla_x f(x)\right\|^2\right)$$

Since $x_1, x_2, \ldots, x_D$ are independent

$$q(x'|x) = \prod_{i=1}^{D} q_i(x_i'|x_i)$$

$$q_i(x_i'|x_i) = Categorical\left(Softmax\left(\frac{1}{2}\nabla_x f(x)_i(x_i' - x_i) - \frac{(x_i' - x_i)^2}{2\alpha}\right)\right)$$

Zhang, et al. A Langevin-like sampler for discrete distributions. ICML 2022.
Benjamin, et al. Enhanced gradient-based MCMC in discrete spaces. TMLR 2022.

Discrete Langevin Proposals

Discrete Langevin Proposals

$$q(x'|x) \propto \exp\left(-\frac{1}{2\alpha}\left\|x' - x - \frac{\alpha}{2}\nabla_x f(x)\right\|^2\right)$$

A fact might interest you

Discrete Langevin dynamics simulates a gradient flow to minimize the KL divergence of the target distribution on a discrete Wasserstein-2 space.

Sun, et al. Discrete Langevin sampler via Wasserstein gradient flow. AISTAST 2023.

Discrete Langevin Proposals

Discrete Langevin Proposals

$$q(x'|x) \propto \exp\left(\frac{f(x') - f(x)}{2} - \frac{||x' - x||^2}{2\alpha}\right)$$

Pitfalls of DLP

From locally-informed to globally-informed

=> the accuracy of the gradient approximation diminishes

Discrete Langevin Proposals

Discrete Langevin Proposals

$$q(x'|x) \propto \exp\left(\frac{f(x') - f(x)}{2} - \frac{||x' - x||^2}{2\alpha}\right)$$

Pitfalls of DLP

From locally-informed to globally-informed

=> the accuracy of the gradient approximation diminishes

$$f(x') - f(x) \approx \nabla_x f(x)^T (x' - x) + \frac{1}{2}(x' - x)^T \nabla_x^2 f(x)(x' - x)$$

increases computational complexity 🤔

Sun, et al. Any-scale Balanced Samplers for Discrete Space. ICLR 2023.

Discrete Langevin Proposals

Discrete Langevin Proposals

$$q(x'|x) \propto \exp\left(\frac{f(x') - f(x)}{2} - \frac{||x' - x||^2}{2\alpha}\right)$$

Pitfalls of DLP

Gradient is ill-defined if natural differentiable extension unavailable

Examples: Facility Location Diversity Models

$$f(S) \coloneqq \sum_{i\in S}\left(\mu_i - \sum_{d=1}^{D} W_{id}\right) + \sum_{d=1}^{D} \max_{i\in S} W_{id}$$

Multilinear Extension

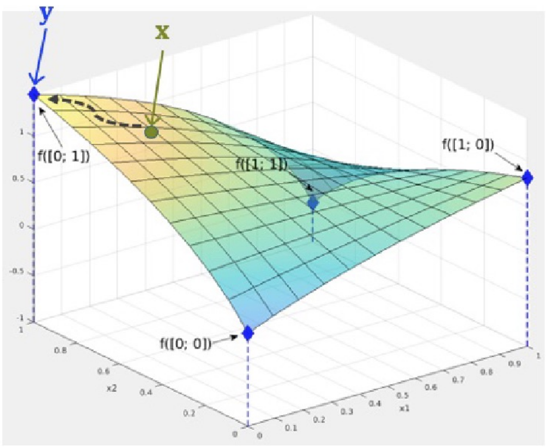
Discrete Langevin Proposals

$$q(x'|x) \propto \exp\left(\frac{f(x') - f(x)}{2} - \frac{||x' - x||^2}{2\alpha}\right)$$

Pitfalls: gradient is ill-defined if natural differentiable extension unavailable

Multilinear Extension

$$f_{mt}(x) := \sum_S f(S) \prod_{i \in S} x_i \prod_{j \notin S} (1 - x_j), \quad x \in [0,1]^D$$



Multilinear Extension

Multilinear Extension Approximations

$$f(x') - f(x) \approx \nabla_x f_{mt}(x)^T (x' - x), \qquad x \in \{0,1\}^D$$

$$\nabla_x f_{mt}(x) \coloneqq \Delta[f](x) \coloneqq (\Delta[f](x)_1, \ldots, \Delta[f](x)_D)$$

$$\Delta[f](x)_i = f(x_{\neg i}) - f(x_i), \qquad \text{if } x \in \{0,1\}^D$$

Xiang, et al. Efficient Informed Proposals for Discrete Distributions via Newton's Series Approximation. AISTAT 2023.

Multilinear Extension

Multilinear Extension Approximations

$$f(x') - f(x) \approx \nabla_x f_{mt}(x)^T (x' - x), \qquad x \in \{0,1\}^D$$

$$\nabla_x f_{mt}(x) \coloneqq \Delta[f](x) \coloneqq (\Delta[f](x)_1, \ldots, \Delta[f](x)_D)$$

$$\Delta[f](x)_i = f(x_{\neg i}) - f(x_i), \qquad \text{if } x \in \{0,1\}^D$$

Connections to Newton's Series Expansion

$$f(x) = \sum_{k=0}^{\infty} \frac{\Delta^k[f](a)}{k!} (x-a)_k$$

$$\Delta^k[f](a) = \sum_{i=0}^{k} \binom{k}{i} (-1)^{k-i} f(x+i) \qquad (x)_k = x(x-1)\cdots(x-k+1)$$

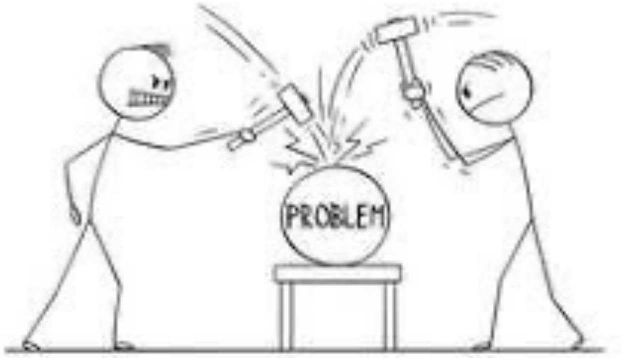
$$f(x') - f(x) \approx \Delta[f](x)^T (x' - x)$$

first-order Newton's series expansion

Xiang, et al. Efficient Informed Proposals for Discrete Distributions via Newton's Series Approximation. AISTAT 2023.

Miscellaneous

Ratio Matching



Discrete EBMs Training

Gradient-Guided Ratio Matching

Ratio Matching

$$\mathcal{L}_{RM}(x;\theta) = \mathbb{E}_{x_{\neg i}\sim U(x_{\neg i})}[\exp(E_\theta(x) - E_\theta(x_{\neg i}))]^2$$

Minimum ratio-matching $\theta^* = \arg\min_\theta \mathcal{L}_{RM}(\theta)$ implies

$$\frac{p_{\theta^*}(x)}{p_{\theta^*}(x_{\neg i})} = \frac{p_{data}(x)}{p_{data}(x_{\neg i})}, \forall i \quad \Rightarrow \quad p_{\theta^*}(x) = p_{data}(x)$$

Liu, et al. Gradient-Guided Importance Sampling for Learning Binary Energy-Based Models. ICLR 2023.

Gradient-Guided Ratio Matching

Ratio Matching

$$\mathcal{L}_{RM}(x;\theta) = \mathbb{E}_{x_{\neg i} \sim U(x_{\neg i})}[\exp(E_\theta(x) - E_\theta(x_{\neg i}))]^2$$

Variance Reduction via Importance Sampling

$$\mathcal{L}_{RM}(x;\theta) = \mathbb{E}_{x_{\neg i} \sim q(x_{\neg i})}\left[\frac{U(x_{\neg i})[\exp(E_\theta(x) - E_\theta(x_{\neg i}))]^2}{q(x_{\neg i})}\right]$$

Liu, et al. Gradient-Guided Importance Sampling for Learning Binary Energy-Based Models. ICLR 2023.

Gradient-Guided Ratio Matching

Ratio Matching

$$\mathcal{L}_{RM}(x;\theta) = \mathbb{E}_{x_{\neg i} \sim U(x_{\neg i})}[\exp(E_\theta(x) - E_\theta(x_{\neg i}))]^2$$

Variance Reduction via Importance Sampling

$$\mathcal{L}_{RM}(x;\theta) = \mathbb{E}_{x_{\neg i} \sim q(x_{\neg i})}\left[\frac{U(x_{\neg i})[\exp(E_\theta(x) - E_\theta(x_{\neg i}))]^2}{q(x_{\neg i})}\right]$$

The optimal proposal is

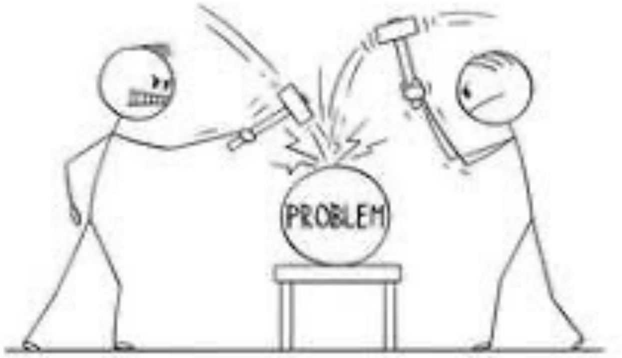
$$q^*(x_{\neg i}) = \frac{[\exp(E_\theta(x) - E_\theta(x_{\neg i}))]^2}{\sum_{d=1}^{D}[\exp(E_\theta(x) - E_\theta(x_{\neg d}))]^2}$$

$$E_\theta(x) - E_\theta(x_{\neg i}) \approx \nabla_x E_\theta(x)^T (x - x_{\neg i})$$

Liu, et al. Gradient-Guided Importance Sampling for Learning Binary Energy-Based Models. ICLR 2023.

Miscellaneous

Concrete Score Matching



Discrete EBMs Training

Concrete Score Matching

Concrete Scores

neighbors of x: $\mathcal{N}(x) = \{x_1, \ldots, x_{n_k}\}$

$$c_{p_{data}}(x; \mathcal{N}) := \left[\frac{p_{data}(x_{n_1}) - p_{data}(x)}{p_{data}(x)}, \ldots, \frac{p_{data}(x_{n_k}) - p_{data}(x)}{p_{data}(x)}\right]^T$$

Meng, et al. Concrete Score Matching: Generalized Score Matching for Discrete Data. NeurIPS 2022.

Concrete Score Matching

Concrete Scores

neighbors of x: $\mathcal{N}(x) = \{x_1, \ldots, x_{n_k}\}$

$$c_{p_{data}}(x; \mathcal{N}) := \left[\frac{p_{data}(x_{n_1}) - p_{data}(x)}{p_{data}(x)}, \ldots, \frac{p_{data}(x_{n_k}) - p_{data}(x)}{p_{data}(x)}\right]^T$$

Concrete Score Matching

$$\mathcal{L}_{\text{CSM}}(\theta) = \sum_x \sum_{i=1}^{|\mathcal{N}(x)|} p_{data}(x) \left(c_\theta(x, \mathcal{N})_i^2 + 2c_\theta(x, \mathcal{N})\right) - \sum_x \sum_i^n 2p_{data}(x_{n_i}) c_\theta(x; \mathcal{N})_i$$

Meng, et al. Concrete Score Matching: Generalized Score Matching for Discrete Data. NeurIPS 2022.

Concrete Score Matching

Concrete Scores

neighbors of x: $\mathcal{N}(x) = \{x_1, \ldots, x_{n_k}\}$

$$c_{p_{data}}(x; \mathcal{N}) := \left[\frac{p_{data}(x_{n_1}) - p_{data}(x)}{p_{data}(x)}, \ldots, \frac{p_{data}(x_{n_k}) - p_{data}(x)}{p_{data}(x)}\right]^T$$

Concrete Score Matching

$$\mathcal{L}_{\mathrm{CSM}}(\theta) = \sum_x \sum_{i=1}^{|\mathcal{N}(x)|} p_{data}(x)\left(c_\theta(x, \mathcal{N})_i^2 + 2c_\theta(x, \mathcal{N})\right) - \sum_x \sum_i^n 2p_{data}(x_{n_i}) c_\theta(x; \mathcal{N})_i$$

Minimum concrete score matching $\theta^* = \arg\min_\theta \mathcal{L}_{CSM}(\theta)$ implies

$$c_{\theta^*}(x, \mathcal{N}) = c_{p_{data}}(x, \mathcal{N}) \forall x \quad \Rightarrow \quad p_{\theta^*}(x) = p_{data}(x)$$

Meng, et al. Concrete Score Matching: Generalized Score Matching for Discrete Data. NeurIPS 2022.

Concrete Score Matching

Concrete Score

$$c_{\theta^*}(x;\mathcal{N}) := \left[\frac{p_{\theta^*}(x_{n_1}) - p_{\theta^*}(x)}{p_{\theta^*}(x)}, \dots, \frac{p_{\theta^*}(x_{n_k}) - p_{\theta^*}(x)}{p_{\theta^*}(x)}\right]^T$$

Inference with Concrete Scores

$$c_{\theta^*}(x;\mathcal{N}) + 1 = \left[\frac{\exp(-E_{\theta^*}(x_{n_1}))}{\exp(-E_{\theta^*}(x))}, \dots, \frac{\exp(-E_{\theta^*}(x_{n_k}))}{\exp(-E_{\theta^*}(x))}\right]^T$$

Meng, et al. Concrete Score Matching: Generalized Score Matching for Discrete Data. NeurIPS 2022.

Concrete Score Matching

Concrete Score

$$c_{\theta^*}(x;\mathcal{N}) := \left[\frac{p_{\theta^*}(x_{n_1}) - p_{\theta^*}(x)}{p_{\theta^*}(x)}, \dots, \frac{p_{\theta^*}(x_{n_k}) - p_{\theta^*}(x)}{p_{\theta^*}(x)}\right]^T$$

Inference with Concrete Scores

$$c_{\theta^*}(x;\mathcal{N}) + 1 = \left[\frac{\exp(-E_{\theta^*}(x_{n_1}))}{\exp(-E_{\theta^*}(x))}, \dots, \frac{\exp(-E_{\theta^*}(x_{n_k}))}{\exp(-E_{\theta^*}(x))}\right]^T$$

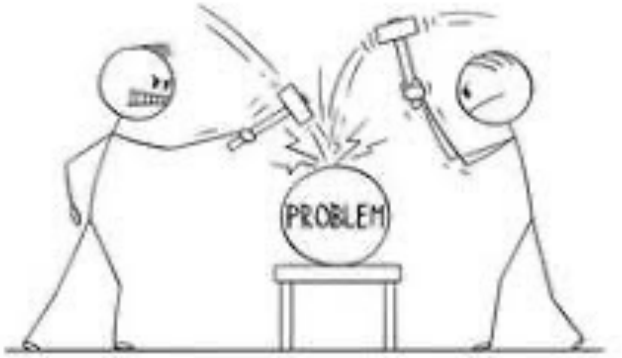
Metropolis Hastings Sampler

$$\min\left\{1, \frac{\exp(-E_{\theta^*}(x'))}{\exp(-E_{\theta^*}(x))}\frac{q(x|x')}{q(x'|x)}\right\}$$

Meng, et al. Concrete Score Matching: Generalized Score Matching for Discrete Data. NeurIPS 2022.

Miscellaneous

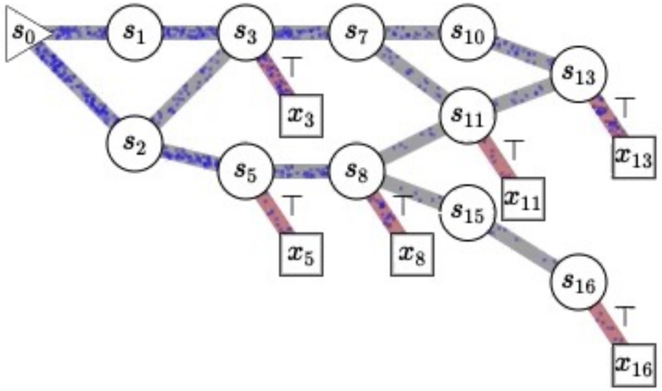
Generative Flow Networks



Discrete EBMs Training

Generative Flow Networks

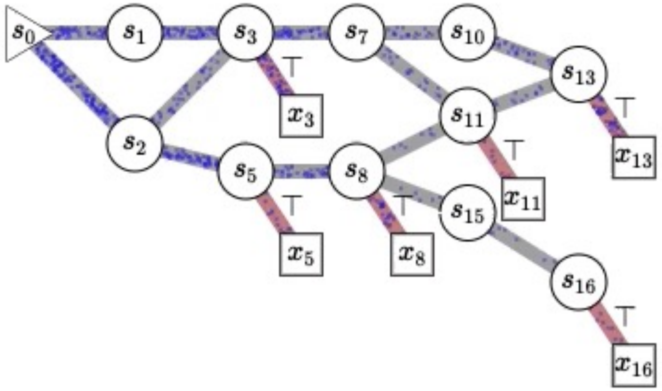
GFlowNet (Learn to Sampling / Amortised MCMC)



Learn to generate **discrete** data x with probability proportional to reward $R(x) > 0$

Generative Flow Networks

GFlowNet (Learn to Sampling / Amortised MCMC)



Learn to generate **discrete** data x with probability proportional to reward $R(x) > 0$

Trajectory balance

$$\min_{\theta} \mathbb{E}_{\tau} \left(\log \frac{Z_{\theta} \prod_{t=1}^{n} p_F(s_t | s_{t-1}; \theta)}{R(x) \prod_{t=1}^{n} p_B(s_{t-1} | s_t; \theta)} \right)^2$$

Optimal θ^* implies that

$$p_{\theta^*}(x) \propto R(x), \qquad x \sim \prod_t p_F(s_t | x_{t-1})$$

Zhang et al. Trajectory Balance: Generative Flow Networks for Discrete Probabilistic Modeling. ICML 2022

Generative Flow Networks

GFlowNet for Discrete EBMs

$$\mathbb{E}_{x\sim p_{data}}\left[\nabla_\phi E_\phi(x)\right] - \mathbb{E}_{x\sim p_\phi}\left[\nabla_\theta E_\phi(x)\right]$$

sample via GFlowNet

$$\text{Step 1:} \quad \min_\theta \mathbb{E}_\tau \left(\log \frac{Z_\theta \prod_{t=1}^{n} p_F(s_t|s_{t-1};\theta)}{\exp\left(-E_\phi(x)\right) \prod_{t=1}^{n} p_B(s_{t-1}|s_t;\theta)}\right)^2$$

$$\text{Step 2:} \quad \min_\phi \mathbb{E}_{x\sim p_{data}}\left[\nabla_\phi E_\phi(x)\right] - \mathbb{E}_{x\sim p_\phi}\left[\nabla_\theta E_\phi(x)\right]$$

sample via $p_F(s_t|s_{t-1})$

Zhang et al. Trajectory Balance: Generative Flow Networks for Discrete Probabilistic Modeling. ICML 2022

References

The family of locally informed proposals

[1] Zanella Giacomo. Informed proposals for local MCMC in discrete spaces. Journal of the American Statistical Association, 2020.
[2] Grathwohl, et al. Oops i took a gradient: Scalable sampling for discrete distributions. ICML, 2021.
[3] Zhang, et al. A Langevin-like sampler for discrete distributions. ICML 2022.
[4] Xiang, et al. Efficient Informed Proposals for Discrete Distributions via Newton's Series Approximation. AISTAT 2023.
[5] Patrick, et al. Plug & Play Directed Evolution of Proteins with Gradient-based Discrete MCMC. Arxiv, 2022.
[6] Benjamin, et al. Enhanced gradient-based MCMC in discrete spaces. TMLR 2022.
[7] Sun, et al. Path auxiliary proposal for mcmc in discrete space. ICLR, 2022.
[8] Sun, et al. Optimal scaling for locally balanced proposals in discrete spaces. NeurIPS 2022.
[9] Sun, et al. Any-scale Balanced Samplers for Discrete Space. ICLR 2023.
[10] Sun, et al. Discrete Langevin sampler via Wasserstein gradient flow. AISTAST 2023.

Raito matching & Concrete score matching

[11] Liu, et al. Gradient-Guided Importance Sampling for Learning Binary Energy-Based Models. ICLR 2023.
[12] Meng, et al. Concrete Score Matching: Generalized Score Matching for Discrete Data. NeurIPS 2022.

GFlowNet

[13] Zhang et al. Trajectory Balance: Generative Flow Networks for Discrete Probabilistic Modeling. ICML 2022

References

Adversarial Training

[14] Dai et al. Learning discrete energy-based models via auxiliary-variable local exploration. NeurIPS 2020

Quasi-Rejection sampling

[15] Eikema et al. An approximate sampler for energy-based models with divergence diagnostics. TMLR 2022

Perturb and MAP

[16] Lazaro-Gredilla et al. Perturb-and-max-product: Sampling and learning in discrete energy-based models. NeurIPS 2021

Thank you!