Energy-Based Modelling for Discrete and Mixed Data via Heat Equations on Structured Spaces

Tobias Schröder *†
Imperial College London

Zijing Ou*‡
Imperial College London

Yingzhen Li
Imperial College London

Andrew Duncan
Imperial College London
The Alan Turing Institute

Abstract

Energy-based models (EBMs) offer a flexible framework for probabilistic modelling across various data domains. However, training EBMs on data in discrete or mixed state spaces poses significant challenges due to the lack of robust and fast sampling methods. In this work, we propose to train discrete EBMs with Energy Discrepancy, a loss function which only requires the evaluation of the energy function at data points and their perturbed counterparts, thus eliminating the need for Markov chain Monte Carlo. We introduce perturbations of the data distribution by simulating a diffusion process on the discrete state space endowed with a graph structure. This allows us to inform the choice of perturbation from the structure of the modelled discrete variable, while the continuous time parameter enables fine-grained control of the perturbation. Empirically, we demonstrate the efficacy of the proposed approaches in a wide range of applications, including the estimation of discrete densities with non-binary vocabulary and binary image modelling. Finally, we train EBMs on tabular data sets with applications in synthetic data generation and calibrated classification.

1 Introduction

Discrete structures are intrinsic to most types of data such as text, graphs, and images. Estimating the data generating distribution p_{data} of discrete data sets with a probabilistic model can contribute greatly to downstream inference and generation tasks, and plays a key role in synthetic data generation of tabular, textual or network data (Raghunathan, 2021). Energy-based models (EBMs) are probabilistic generative models of the form $p_{\text{ebm}} \propto \exp(-U)$, where the flexible choice of the energy function U allows great control in the modelling of different data structures. However, energy-based models are, by definition, unnormalised models and notoriously difficult to train due to the intractability of their normalisation, especially in discrete spaces.

Energy-based models are typically trained with the contrastive divergence (CD) algorithm (Hinton, 2002) which performs approximate maximum likelihood estimation by approximating the gradient of the log-likelihood with Markov Chain Monte Carlo (MCMC) techniques. This method motivated rich research results on sampling from discrete distributions to enable fast and accurate estimation of energy-based models (Zanella, 2020; Grathwohl et al., 2021; Zhang et al., 2022b; Sun et al., 2022b,a, 2023a). However, the training of energy-based models with CD remains challenging as it relies on

*Equal contributions
†Correspondence to: t.schroeder21@imperial.ac.uk and z.ou22@imperial.ac.uk
‡Code: https://github.com/J-zin/discrete-energy-discrepancy

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-2516

sufficiently fast mixing of Markov chains. Since accurate sampling from the EBM typically cannot be achieved, contrastive divergence lacks theoretical guarantees (Carreira-Perpinan & Hinton, 2005) and leads to biased estimates of the energy landscape (Nijkamp et al., 2019). For mixed data types, energy-based models have only been applied to combinations of numerical features with a single categorical label (Grathwohl et al., 2020).

The recently introduced Energy Discrepancy (ED) (Schröder et al., 2023) is a new type of contrastive loss functional that, by definition, depends on neither gradients nor MCMC methods. Instead, the definition of ED only requires the evaluation of the energy function on positive and contrasting, negative samples which are generated by perturbing the data distribution. However, the work in Schröder et al. (2023) is currently limited to Gaussian perturbations on continuous spaces and does not explores strategies to choose perturbations on discrete spaces, especially when these discrete spaces exhibit some additional structure.

In this work, we propose a framework to train energy-based models with energy discrepancy on discrete data, making the following contributions: 1) We explore a method to define discrete diffusion processes on structured discrete spaces through a heat equation on the underlying graph and investigate the effect of geometry and time parameter on the diffusion. 2) Based on the discrete diffusion process, we extend energy discrepancy to discrete spaces in a systematic way, thus introducing a MCMC-free method for the training of energy-based models that requires little tuning. 3) We extend our methodology to mixed state spaces and establish to the best of our knowledge the first robust training method of energy-based models on tabular data sets. We demonstrate promising performance on downstream tasks like synthetic data generation and calibrated prediction, thus unlocking a new tool for generative modelling on tabular data.

2 Preliminaries

Energy-based models (EBMs) are a parametric family of distributions p_θ defined as

$$p_\theta(\mathbf{x}) = \frac{\exp(-U_\theta(\mathbf{x}))}{Z_\theta}, \quad Z_\theta = \sum\nolimits_{\mathbf{x}\in\mathcal{X}} \exp(-U_\theta(\mathbf{x})), \tag{1}$$

where U_θ is the energy function parameterised by θ and Z_θ denotes the normalisation constant. Given a set of *i.i.d.* samples $\{\mathbf{x}^i\}_{i=1}^N$ from an unknown data distribution $p_{\text{data}}(\mathbf{x})$ we aim to learn an approximation $p_\theta(\mathbf{x})$ of $p_{\text{data}}(\mathbf{x})$. The *de facto* standard approach for finding such θ is to minimise the negative log-likelihood of p_θ under the data distribution via gradient decent

$$-\nabla_\theta \mathbb{E}_{p_{\text{data}}(\mathbf{x})}[\log p_\theta(\mathbf{x})] = \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[\nabla_\theta U_\theta(\mathbf{x})] - \mathbb{E}_{\mathbf{x}_-\sim p_\theta}[\nabla_\theta U_\theta(\mathbf{x}_-)]. \tag{2}$$

The intuition behind this update is to decrease the energy of positive data samples $\mathbf{x} \sim p_{\text{data}}(\mathbf{x})$ and to increase the energy of negative samples $\mathbf{x}_- \sim p_\theta(\mathbf{x})$. However, the exact computation of gradient in (2) is known to be NP-hard in general (Jerrum & Sinclair, 1993) and quickly becomes prohibitive even on relatively simple data sets. Consequently, existing approaches resort to sampling from the model p_θ to approximate the gradient of log-likelihood via Monte Carlo estimation. In discrete settings, the most popular sampling methods include the locally informed sampler (Zanella, 2020), Gibbs with gradients (GwG) (Grathwohl et al., 2021), discrete Langevin (Zhang et al., 2022b), and generative flow networks (GFlowNet) (Zhang et al., 2022a). Despite their established success in discrete energy-based modelling, these methods necessitate a trade-off that hampers scalability: running the sampler for an extended duration rapidly increases the cost of maximum likelihood training, while shorter sampler runs yield inaccurate approximations of the likelihood gradient and introduce biases into the learned energy.

Energy Discrepancy (Schröder et al., 2023) is a recently proposed method to train energy-based models without the need for an extensive sampling process. Instead, it constructs negative samples by perturbing the data, thus bypassing the sampling step while still yielding a valid training objective. To elucidate, the energy discrepancy is formally defined as follows:

Definition 1 (Energy Discrepancy)**.** *Let p_{data} be a positive density on a measure space $(\mathcal{X}, \mathrm{d}\mathbf{x})$*[4] *and let $q(\mathbf{y}|\mathbf{x})$ be a conditional probability density. Define the* contrastive potential *induced by q as*[5]

$$U_q(\mathbf{y}) := -\log \sum_{\mathbf{x}'\in\mathcal{X}} q(\mathbf{y}|\mathbf{x}') \exp(-U(\mathbf{x}')) \tag{3}$$

The energy discrepancy *between* p_{data} *and* U *induced by* q *is defined as*

$$\text{ED}_q(p_{\text{data}}, U) := \mathbb{E}_{p_{\text{data}}(\mathbf{x})}[U(\mathbf{x})] - \mathbb{E}_{p_{\text{data}}(\mathbf{x})}\mathbb{E}_{q(\mathbf{y}|\mathbf{x})}[U_q(\mathbf{y})]. \tag{4}$$

We will refer to q as the *perturbation*. The validity of this loss functional was proven in Schröder et al. (2023) in large generality: In particular, it is sufficient for $U^* = \operatorname{argmin} \text{ED}_q(p_{\text{data}}, U) \Leftrightarrow \exp(-U^*) \propto p_{\text{data}}$ that any two points $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ are q-equivalent, i.e. there exists a chain of states $(\mathbf{z}^i)_{i=1}^T \in \mathcal{X}$ with $\mathbf{z}^1 = \mathbf{x}, \mathbf{z}^T = \mathbf{y}$ such that $q(\mathbf{z}^{i+1}|\mathbf{z}^i) > 0$ for all $i = 1, \dots, T-1$.

Energy discrepancy can also be understood from seeing it as a type of Kullback-Leibler divergence. Specifically, the loss function defined in (4) is equivalent to the expected Kullback-Leibler divergence

$$\operatorname*{argmin}_U \text{ED}_q(p_{\text{data}}, U) \Leftrightarrow \operatorname*{argmin}_U \mathbb{E}_{p_{\text{data}}(\mathbf{x})}\mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\text{KL}(p_{\text{data}}(\cdot|\mathbf{y}), p_{\text{ebm}}(\cdot|\mathbf{y}))\right] \tag{5}$$

where $p_\bullet(\mathbf{x}|\mathbf{y}) \propto p_\bullet(\cdot)q(\mathbf{y}|\cdot)$. This relates energy discrepancy to diffusion recovery likelihood objectives (Gao et al., 2021) and Kullback-Leibler contractions (Lyu, 2011). Energy discrepancy has demonstrated notable effectiveness in training EBMs in continuous spaces (Schröder et al., 2023). In the next section, we show how the loss can be defined in discrete spaces.

3 Energy Discrepancies for Discrete Data

For this work we will first consider a state space for the data distribution that can be written as the product of d discrete variables with S_k classes each, i.e. $\mathcal{X} = \bigotimes_{k=1}^d \{1, \dots, S_k\}$. Examples for spaces of this type are the categorical entries of a data table for which d denotes the number of features, or binary image data sets for which we typically write $\mathcal{X} = \{0, 1\}^d$. To define energy discrepancy on such spaces we need to specify a perturbation process under the following considerations: 1) The negative samples obtained through q are informative for training the EBM when only finite amounts of data are available. 2) The contrastive potential $U_q(\mathbf{y})$ has a numerically tractable approximation.

Let us consider one component $\mathcal{X} = \{1, \dots, S\}$, only. Inspired from previous works on diffusion modelling for discrete data (Campbell et al., 2022; Sun et al., 2023b; Lou et al., 2024; Campbell et al., 2024) we model the perturbation as a continuous time Markov chain (CTMC) with transition probability

$$q_t(y = b|x = a) = \exp\left(tR\right)_{ba}, \quad a, b \in \{1, 2, \dots, S\}$$

where $R \in \mathbb{R}^{S \times S}$ is the so-called rate matrix which satisfies $R_{bb} = -\sum_{a \neq b}^S R_{ba}$ and $\exp(tR)$ is the matrix exponential. For a given rate matrix R, this approach then leaves us with a single tunable hyperparameter t characterising the magnitude of perturbation applied. We first analyse how the choice of rate matrix and time parameter affect the statistical properties of the energy discrepancy loss. In fact, under weak conditions, the energy discrepancy loss converges to maximum likelihood estimation for $t \to \infty$, thus achieving the same loss function implemented by contrastive divergence:

Theorem 1. *Let* $q_t(\cdot|x)$ *be a Markov transition density defined by the rate matrix* R *with eigenvalues* $0 = \lambda_1(R) \geq \lambda_2(R) \geq \cdots \geq \lambda_S(R)$ *and uniform stationary distribution. Then, there exists a constant* z_t independent *of* θ *such that energy-discrepancy converges to the maximum-likelihood loss*

$$|\text{ED}_{q_t}(p_{\text{data}}, U_\theta) - \mathcal{L}_{\text{MLE}}(\theta) - z_t| \leq \sqrt{S} \exp(-|\lambda_2(R)|t) \, \text{KL}(p_{\text{data}} \parallel p_\theta) \, .$$

with the loss of maximum-likelihood estimation $\mathcal{L}_{\text{MLE}}(\theta) := -\mathbb{E}_{p_{\text{data}}(\mathbf{x})}\big[\log p_\theta(\mathbf{x})\big]$.

Here, z_t is a constant independent of θ, so the optimisation landscapes of energy discrepancy estimation and maximum likelihood estimation in θ align at an exponential rate, except for a shift by z_t which does not affect the optimisation. This result improves the linear convergence rate in Schröder et al. (2023) and relates it to the *spectral gap* $|\lambda_2(R)|$ of the rate matrix. Such a result is meaningful as the maximum-likelihood estimator is generally statistically preferable with better sample efficiency, and Theorem 1 suggests that energy discrepancy estimation can approximate maximum likelihood estimation without resorting to MCMC like in classical EBM training methods. The proof is given in Appendix A.1.

[4] On discrete spaces $\mathrm{d}\mathbf{x}$ is assumed to be a counting measure. On continuous spaces $\mathcal{X}$, the appearing sums and expectations turn into integrals with respect to the Lebesgue measure

[5] With a slight abuse of notations, we represent the contrastive potential induced by distribution q as U_q and denote the energy function as U with or without the subscript θ.

https://doi.org/10.52202/079017-2516

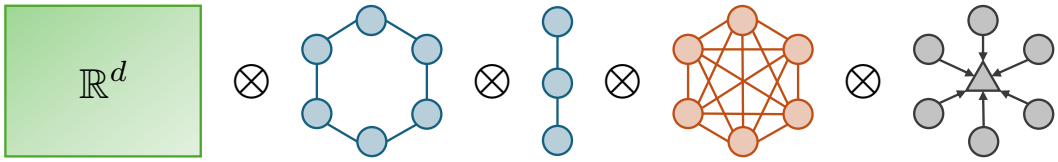


Figure 1: Visualisation of a typical state space of a tabular dataset: Numerical entries taking values in $\mathbb{R}^d$, cyclical categorical entries (e.g. season), ordinal categorical entries (e.g. age), unstructured categorical entries, and variables with an absorbing state associated with masking the entry.

3.1 Heat Equation in Structured Discrete Spaces

In principle, Theorem 1 establishes that energy discrepancy converges to the loss of maximum likelihood estimation in the limit $t \to \infty$ for any choice of rate matrix with spectral gap. In practice, however, large perturbations of data can produce high-variance parameter gradients and provide little training signal. Instead, it is desirable to construct perturbations that allow a fine-grained trade-off between the statistical properties of the loss function and the variance of the gradients. For this reason, we investigate the perturbation for small t which, as we will see, can be informed by the assumed *graph structure* of the underlying discrete space.

The infinitesimal perturbations of the CTMC are characterised by the rate matrix. Notice that for small t, the Euler discretisation of the heat equation yields

$$q_t(y = b|x = a) \approx \delta(b, a) + tR_{ba}\,.$$

with $\delta(b, a) = 1 \Leftrightarrow a = b$ and zero otherwise. To model the relationship between values $a, b \in \{1, \dots, S\}$ we endow the space with a graph structure with adjacency matrix A and out degree matrix D_{out} and model the rate matrix as the graph Laplacian $R := (A - D_{\text{out}})$. By definition, the columns of the graph Laplacian matrix sum to zero $\sum_{b=1}^{S} R_{ba} = 0$. The smallest possible perturbation is then characterised as the transition to an adjacent neighbour. The characterisation of the CTMC in terms of the graph Laplacian is implicitly assumed in previous work. Campbell et al. (2022) describe a diffusion via a uniform perturbation which corresponds to a fully connected graph and Lou et al. (2024) describe the rate matrix associated to a star graph with absorbing (masking) state $\square$:

$$R^{\text{unif}}_{ba} = 1 - S\,\delta(b, a)\,, \qquad R^{\text{mask}}_{ba} = \delta(b, \square) - \delta(b, a)\,.$$

For a visualisation, see Section 3.1. In addition to these fairly unstructured rate matrices we model state spaces with a cyclical or ordinal structure:

$$R^{\text{cyc}} = \begin{pmatrix} -2 & 1 & 0 & \cdots & 0 & 1 \\ 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & 0 & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & & 0 & 1 & -2 & 1 \\ 1 & 0 & \cdots & 0 & 1 & -2 \end{pmatrix} \qquad R^{\text{ord}} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & 0 & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & & 0 & 1 & -2 & 1 \\ 0 & 0 & \cdots & 0 & 1 & -1 \end{pmatrix}.$$

We restrict ourselves to uniform, cyclical, and ordinal structures as these structures can typically be trivially inferred from the type of data modelled. For example, periodically changing quantities (e.g. seasons) would display a cyclical structure and ordered information like age an ordinal structure. It is possible, however, to extend our framework to arbitrary graphical structures of the state space as long as eigenvalue decompositions of the graph Laplacian are feasible.

Since the Gaussian perturbation on Euclidean space used in Schröder et al. (2023) is also the solution of a heat equation, these choices allow us to model the perturbation on a vector of mixed entries including numerical values, unstructured categorical values, and structured categorical values with a single differential equation

$$\partial_t q_t(\cdot|\mathbf{x}) = \left(\Delta^{\text{num}} \otimes R^{\text{cyc}} \otimes R^{\text{ord}} \otimes R^{\text{unif}} \otimes R^{\text{abs}}\right) q_t(\cdot|\mathbf{x})\,, \quad q_0(\cdot|\mathbf{x}) = \delta(\cdot, \mathbf{x}) \tag{6}$$

where Δ^{num} denotes the standard Laplace operator $\sum_{k=1}^{d} \partial^2 / \partial^2_{\mathbf{x}^{\text{num}}}$ and the product $\otimes$ denotes that each operator acts on the corresponding component of the state space.

4 Estimating the Energy Discrepancy Loss

We now discuss how discrete energy discrepancy can be estimated. We will typically assume that each dimension of the data point is perturbed independently, i.e. the perturbation $q(\mathbf{y}|\mathbf{x})$ is modelled as the product of component-wise perturbations. On Euclidean data, we resort to the implementation in Schröder et al. (2023) and obtain perturbed samples by adding isotropic Gaussian noise to the samples. We are now left with the heat equation on discrete space.

4.1 Solving the Heat Equation

In the case of the uniform Laplacian $R_{\text{unif}} = \mathbf{1}_S \mathbf{1}_S^T - S\text{id}$, with $\mathbf{1}_S$ denoting an all ones vector of length S, the heat equation has the closed form solution

$$q_t(y|x=a) = e^{-St}\delta(y,a) + \frac{1-e^{-St}}{S}\sum_{b=1}^{S}\delta(y,b)\,. \tag{7}$$

Practically speaking, this perturbation remains in its state with probability e^{-St} and samples uniformly from the state space otherwise. The case of the cyclical and ordinal structure is more delicate. We first note that the heat equation can be solved in terms of its eigenvalue expansion $\exp(Rt) = \mathbf{V}\exp(\Lambda t)\mathbf{V}^*$, where Λ is the matrix with the eigenvalues λ_p along its diagonal and $\mathbf{V}$ is a matrix of orthogonal eigenvectors with each column containing the corresponding eigenvector $\mathbf{v}_p$. The perturbation for R^{cyc} and R^{ord} can then be computed by means of a discrete Fourier transform:

Proposition 1. *Assume the density $q_t(b|a) := q_t(y=b|x=a)$ is defined by the rate matrices R^{cyc} or R^{ord}. The transition density for all $a, b \in \{1, 2, \dots, S\}$ is given by*

$$q_t^{\text{cyc}}(b|a) = \frac{1}{S}\sum_{p=1}^{S}\exp(2\pi\mathrm{i}b\omega_p^{\text{cyc}})\,\exp\left((2\cos(2\pi\omega_p^{\text{cyc}})-2)t\right)\,\exp(-2\pi\mathrm{i}a\omega_p^{\text{cyc}}) \tag{8}$$

$$q_t^{\text{ord}}(b|a) = \frac{2}{S}\sum_{p=1}^{S}\frac{1}{z_p}\,\cos((2b-1)\pi\omega_p^{\text{ord}})\,\exp\left((2\cos(2\pi\omega_p^{\text{ord}})-2)t\right)\,\cos((2a-1)\pi\omega_p^{\text{ord}}) \tag{9}$$

where $\omega_p^{\text{cyc}} = (p-1)/S$ and $\omega_p^{\text{ord}} = (p-1)/2S$, respectively, and $z_p = (2, 1, \dots, 1)$.

For the derivation, see Appendix A.2. Due to this result, the heat equation can be efficiently solved in parallel without requiring any sequential operations like multiple Euler steps. In addition, the transition matrices can be computed and saved in advance, thus reducing the computational complexity to the matrix multiplication with a batch of one-hot encoded data points.

Gaussian limit and choice of time parameter. For tabular data sets the cardinality S changes between different dimensions which raises the question how t should be scaled with S. To answer this question we observe the following scaling limit of the perturbation:

Theorem 2 (Scaling limit)**.** *Let $y_t \sim q_t(\cdot|x=\mu S)$ with $\mu \in \{1/S, 2/S, \dots, 1\}$, where q_t is either the transition density of the cyclical or ordinal perturbation. Let $\varphi : \mathbb{R} \to (0,1]$, where for all $n \in \mathbb{Z}$ and $x \in (0,1]$ $\varphi^{\text{cyc}}(n+x) = x$ and $\varphi^{\text{ord}}(2n+x) = x$, $\varphi^{\text{ord}}(2n+1+x) = -x$. Then,*

$$y_{S^2t}/S \xrightarrow{S\to\infty} \varphi(\xi_t) \quad \textit{with} \quad \xi_t \sim \mathcal{N}(\mu, 2t)\,.$$

Consequently, under the rescaling of time and space prescribed, the perturbation behaves independently of the state space size like a Gaussian with variance $2t$ that is reflected or periodically continued at the boundary states of $(0,1]$. The phenomenon is visualised in Figure 5. Based on this scaling limit we typically choose a quadratic rule $t = S^2 t_{\text{base}}$. Alternatively, we may choose a linear rule $t = St_{\text{base}}$ in which case the limit becomes a regular Gaussian on $\mathbb{R}_+$, thus recovering the Euclidean case from Schröder et al. (2023). The theorem is proven in Appendix A.3.

As a byproduct of this result we can also approximate the perturbation with discretised rescaled samples from a standard normal distribution and applying either periodic or reflecting mappings on perturbed states outside the domain. This may be computationally favourable for spaces of the form $\{1, \dots, S\}^d$ where the vocabulary size S and dimension of the state space d grow very large.

https://doi.org/10.52202/079017-2516

Localisation to random grid. For unstructured categorical variables the uniform perturbation may introduce too much noise to inform the EBM about the correlations in the data set. In these cases, it can be beneficial to sample a random dimension $k \in \{1, \dots, d\}$ and apply a larger perturbation in this dimension, only. This effectively means to replace the product of categorical distributions with a mixture perturbation

$$q_t(\mathbf{y}|\mathbf{x}) = \prod_{k=1}^{d} q_t(y_k|x_k) \to \frac{1}{d} \sum_{k=1}^{d} q_t(y_k|x_k)\,.$$

In our experiments we only consider the case of perturbing the randomly chosen dimension uniformly. We call this grid perturbation due to connections with concrete score matching (Meng et al., 2022). The resulting loss can be understood as a variation of pseudo-likelihood estimation.

Special case of binary state space. In the special case of $\mathcal{X} = \{0, 1\}^d$, the structures of the cyclical, ordinal, and uniform graph coincide, and the perturbation $q_t(\mathbf{y}|\mathbf{x})$ becomes the product of identical Bernoulli distributions with parameter $\varepsilon = 0.5(1 - \mathrm{e}^{-2t})$. We also explore the grid perturbation which assumes that a dimension is selected at random and the entry is flipped deterministically from zero to one or one to zero. For details, see Appendix B.1.

4.2 Estimation of the Contrastive Potential

The final challenge in turning energy discrepancy into a practical loss function lies in the estimation of the contrastive potential U_q. We use the fact that for a symmetric rate matrix R, the induced perturbation is symmetric as well, i.e. $q_t(\mathbf{y}|\mathbf{x}) = q_t(\mathbf{x}|\mathbf{y})$. Thus, we first write the contrastive potential as an expectation $U_q(\mathbf{y}) = -\log \sum_{\mathbf{x}\in\mathcal{X}} \exp(-U(\mathbf{x}))q(\mathbf{y}|\mathbf{x}) = -\log \mathbb{E}_{q(\mathbf{x}|\mathbf{y})}\left[\exp(-U(\mathbf{x}))\right]$ and subsequently approximate the energy discrepancy loss as in Schröder et al. (2023) as $\mathcal{L}_{q,M,w}(U) := \frac{1}{N}\sum_{i=1}^{N} \log\left(w + \sum_{j=1}^{M} \exp(U(\mathbf{x}^i) - U(\mathbf{x}_-^{i,j})\right)$ with $\mathbf{x}^i \sim p_{\mathrm{data}}$, $\mathbf{y}^i \sim q_t(\cdot|\mathbf{x}^i)$, and $\mathbf{x}_-^{i,j} \sim q_t(\cdot|\mathbf{y}^i)$. The details of the training procedure are provided in Appendix C.

5 Related Work

Contrastive loss functions. Our work is based on energy discrepancies first introduced in (Schröder et al., 2023). Energy discrepancies are equivalent to certain types of KL contraction divergences whose theory was studied in Lyu (2011), however, without proposing a training algorithm for EBM's. On Euclidean data, ED is related to diffusion recovery-likelihood (Gao et al., 2021) which uses a CD-type training algorithm. For a masking perturbation, ED estimation can be understood as a Monte-Carlo approximation of pseudo-likelihood (Besag, 1975). Furthermore, the structure of the stabilised energy discrepancy loss shares similarities with other contrastive losses such as Ceylan & Gutmann (2018); Gutmann & Hyvärinen (2010); Oord et al. (2018); Foster et al. (2020) due to their close connection to the Kullback-Leibler divergence.

Discrete diffusion models. We extend the continuous time Markov chain framework introduced and developed in Campbell et al. (2022, 2024); Lou et al. (2024) and provides a geometric interpretation thereof. Similar to us, Kotelnikov et al. (2023) defines a flow on mixed state spaces as the product of a Gaussian and a categorical flow, utilising multinomial flows (Hoogeboom et al., 2021). Our work has connections to concrete score matching (Meng et al., 2022) through the usage of neighbourhood structures to define a replacement of the continuous score function.

Contrastive divergence and sampling. Contrastive divergence (CD) is commonly utilised for training energy-based models in continuous spaces with Langevin dynamics (Xie et al., 2016, 2018, 2022; Du et al., 2021; Xiao et al., 2020). In discrete spaces, EBM training heavily relies on CD methods as well, which is a major driver for the development of discrete sampling strategies. The standard Gibbs method was improved by Zanella (2020) through locally informed proposals. This method was extended to include gradient information (Grathwohl et al., 2021) to drastically reduce the computational complexity of flipping bits in several places (Sun et al., 2022b; Emami et al., 2023; Sun et al., 2022a). Moreover, a discrete version of Langevin sampling was introduced based on this idea (Zhang et al., 2022b; Rhodes & Gutmann, 2022; Sun et al., 2023a). Consequently, most current implementations of contrastive divergence use multiple steps of a gradient-based discrete sampler. Alternatively, EBMs can be trained using generative flow networks which learn a Markov chain that construct data optimising the energy as reward function (Zhang et al., 2022a).

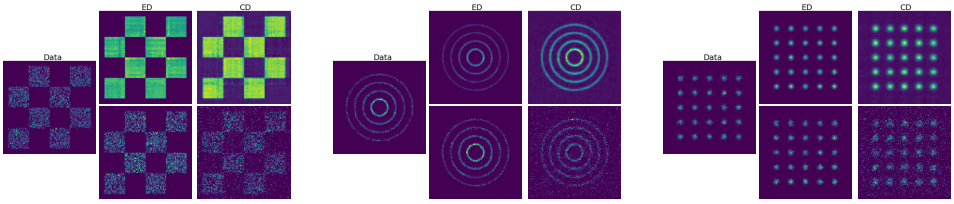


Figure 2: Comparison of energy discrepancy and contrastive divergence on the dataset with 16 dimensions and 5 states. Rows 1 and 2 show the estimated density and synthesised samples, respectively.

Other training methods and applications of EBMs for discrete and mixed data. A sampling-free approach for training binary discrete EBMs is ratio matching (Hyvärinen, 2007; Lyu, 2009; Liu et al., 2023). Dai et al. (2020) propose to apply variational approaches to train discrete EBMs instead of MCMC. Eikema et al. (2022) replace the widely-used Gibbs algorithms with quasi-rejection sampling to trade off the efficiency and accuracy of the sampling procedure. The perturb-and-map (Papandreou & Yuille, 2011) is also recently utilised to sample and learn in discrete EBMs (Lazaro-Gredilla et al., 2021). Tran et al. (2011) introduce mixed-variate restricted Boltzmann machines for energy-based modelling on mixed state spaces. Deep architectures, on the other hand, have been mostly limited to a single categorical target variable which is modelled via a classifier (Grathwohl et al., 2020). Moreover, Ou et al. (2022) apply discrete EBMs on set function learning, in which the discrete energy function is learned with approximate marginal inference (Domke, 2013).

6 Experiments

To evaluate our proposed approach, we conduct experiments across diverse scenarios: i) estimating probability distributions on discrete data; ii) handling mixed-state features in tabular data; and iii) modelling binary images. We also explore Ising model training and graph generation in binary spaces, but leave the detailed evaluation of these in Appendix D.

6.1 Discrete Density Estimation

We first demonstrate the effectiveness of energy discrepancy on density estimation using synthetic discrete data. Following Dai et al. (2020), we initially generate 2D floating-point data from several two-dimensional distributions. Each dimension of the data is then converted into a 16-bit Gray code, resulting in a dataset with 32 dimensions and 2 states. To construct datasets beyond binary cases, we follow Zhang et al. (2024) and transform each dimension into 8-bit 5-base code and 6-bit decimal code. This process creates two additional datasets: one with 16 dimensions and 5 states, and another with 12 dimensions and 10 states. The experimental details are given in Appendix D.1.

Figure 2 illustrates the estimated energies as well as samples that are synthesised with Gibbs sampling for energy discrepancy (ED) and contrastive divergence (CD) on the dataset with 16 dimensions and 5 states. It can be seen that ED excels at capturing the multi-modal nature of the distribution, consistently learning sharper energy landscape in the data support compared to CD. This coincides with the previous observations in continuous spaces (Schröder et al., 2023), suggesting ED's advantage in handling complex data structures. For more results of additional datasets with 5 and 10 states, we deferred them to Figures 7 and 8, respectively.

For binary cases with 2 states, we compare our approaches to three baselines: PCD (Tieleman, 2008), ALOE+ (Dai et al., 2020), and EB-GFN (Zhang et al., 2022a). In Tables 3 and 4, we quantitatively evaluate different methods by evaluating the negative log-likelihood (NLL) and the exponential Hamming MMD (Gretton et al., 2012), respectively. We observe that energy discrepancy outperforms the baseline methods in most settings, but without relying on MCMC simulations (as in PCD) or the training of additional variational networks (as in ALOE and EB-GFN). This performance gain is likely explained by the good theoretical guarantees of energy discrepancy for well-posed estimation tasks. In contrast, the baselines introduce biases due to their reliance on variational proposals and short-run MCMC sampling that may not have converged.

https://doi.org/10.52202/079017-2516

Table 1: Results on real-world datasets.

Methods	**Adult**	**Bank**	**Cardio**	**Churn**	**Mushroom**	**Beijing**	**Avg. Rank**
	AUC ↑	AUC ↑	AUC ↑	AUC ↑	AUC ↑	RMSE ↓	–
Real	.927±.000	.935±.002	.834±.001	.819±.001	1.00±.000	.423±.003	–
CTGAN	.861±.005	.774±.006	.787±.002	.792±.003	.781±.007	1.01±.038	6.33
TVAE	.873±.001	.868±.002	.676±.009	.793±.006	.999±.000	1.05±.012	5.17
TabCD	.619±.026	.604±.021	.765±.008	.584±.021	.561±.048	1.06±.037	8.83
TabDDPM	.910±.001	.922±.001	.801±.001	.806±.007	.999±.000	.556±.005	1.5
TabED-Uni	.884±.003	.842±.013	.786±.002	.810±.008	.998±.001	1.04±.013	3.83
TabED-Grid	.833±.003	.831±.004	.791±.001	.803±.007	.985±.005	.978±.015	4.83
TabED-Cyc	.831±.005	.823±.007	.796±.001	.807±.007	.971±.004	1.01±.024	4.83
TabED-Ord	.853±.005	.845±.004	.792±.002	.806±.004	.926±.010	1.02±.017	4.83
TabED-Str	.879±.004	.819±.001	-	-	-	.978±.012	3.67

6.2 Tabular Data Synthesising

In this experiment, we assess our methods on synthesising tabular data, which presents a challenge due to its mix of numerical and categorical features, making it more difficult to model compared to conventional data formats. To demonstrate the efficacy of energy discrepancies, we first conduct experiments on synthetic examples before proceeding to real-world tabular data. Additional details regarding the experimental setup are deferred to Appendix D.2.

Synthetic Dataset. We first showcase the effectiveness of our methods on mixed data types by learning EBMs on a synthetic ring dataset. The dataset consists of four columns, with the first two columns indicating numerical coordinates of data points. The third column categorizes data points into four circles whereas the last column specifies the 16 colours each data point could be classified into. Therefore, each row in the tabular contains 2 numerical features and 2 categorical features.

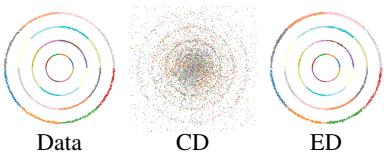


Figure 3: Comparison of the energy discrepancy and contrastive divergence on the synthetic tabular datasets.

To train an EBM on a dataset comprising mixed types of data, we employ either contrastive divergence or energy discrepancy. For CD, we adopt a strategy involving a replay buffer in conjunction with a short-run MCMC using 20 steps. Specifically, we utilise Langevin dynamics and Gibbs sampling for numerical and categorical features, respectively. In the case of ED, we perturb the numerical features with a Gaussian perturbation and the categorical features with grid perturbation. Figure 3 illustrates the results of synthesised samples generated from the learned energy using Gibbs sampling. These findings align with those depicted in Figure 2, where CD struggles to capture a faithful energy landscape, leading to synthesized samples potentially lying outside the data distribution support. Instead, by leveraging a combination of perturbation techniques tailored to the data types present, ED offers a more robust and reliable framework for training EBMs in mixed state spaces.

Real-world Dataset. We then evaluate our methods by benchmarking them against various baselines across 6 real-world datasets. Following Xu et al. (2019), we first split the real datasets into training and testing sets. The generative models are then learned on the real training set, from which synthetic samples of equal size are generated. This synthetic dataset is subsequently used to train a classification/regression XGBoost model, which is evaluated using the real test set.

We compare the performance, as measured by the AUC score for classification tasks and RMSE for regression tasks, against CTGAN, TVAE, (Xu et al., 2019) and TabDDPM (Kotelnikov et al., 2023) baselines which utilise generative adversarial networks, variational autoencoders, and denoising diffusion probabilistic models, respectively. The results are reported in Table 1. Here, TabED-Str refers to an ED loss for which the perturbation was chosen with prior knowledge about the structure of the modelled feature, i.e. ordinal and cyclical features were hand-picked. We do not report results for TabED-Str on the Cardio, Churn, and Mushroom datasets, since the state spaces only consist

Table 2: Experimental results for discrete image modelling. We report the negative log-likelihood (NLL) on the test set for different models. The results of Gibbs, GWG, and DULA are taken from Zhang et al. (2022b), and the result of EB-GFN is from Zhang et al. (2022a).

Dataset \ Method	Gibbs	GWG	EB-GFN	DULA	ED-Bern	ED-Grid
Static MNIST	117.17	**80.01**	102.43	80.71	96.11	90.61
Dynamic MNIST	121.19	**80.51**	105.75	81.29	97.12	90.19
Omniglot	142.06	94.72	112.59	145.68	97.57	**93.94**

of unstructured features. To compute the average ranking we use the rank of TabED-Uni on these datasets since on unstructured features TabED-Uni and TabED-Str coincide.

The variants of ED show promising results on diverse datasets, thus demonstrating the suitability of ED for EBM training on mixed-state spaces. While TabDDPM outperforms the other approaches, TabED shows comparable performance to the CTGAN and TVAE baselines and outperforms both in average ranking. Furthermore, the contrastive divergence approach performs poorly which highlights its limitations in accurately modelling distributions on mixed state spaces. Surprisingly, the unstructured perturbation TabED-Uni performs slightly better than the structured approaches. This may partially be attributed to the fact that the state spaces of the discrete features are relatively small. Consequently, the uniform perturbation might be a good approximation of maximum likelihood estimation in agreement with Theorem 1, while not producing high-variance gradients on these specific datasets.

Improving Calibration. Despite the improving accuracy of neural-network-based classifiers in recent years, they are also becoming increasingly recognised for their tendency to exhibit poor calibration due to over-confident outputs (Guo et al., 2017; Mukhoti et al., 2020). Since energy-based model on mixed state spaces can capture the likelihood of tuples of features and target labels, they implicitely quantify the confidence in a prediction and can be adapted into classifiers with better calibration than deterministic methods. This opens up a new avenue for applying EBMs in deterministic tabular data modelling methods.

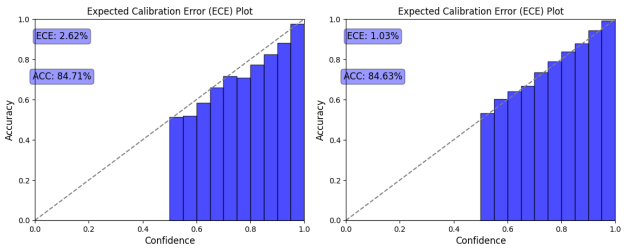


Figure 4: Calibration results comparison between the baseline (left) and energy discrepancy (right) on the adult dataset.

Let y and $\mathbf{x}$ be the target label and the rest features in the tabular data, an EBM $U_\theta(\mathbf{x}, y)$ learned on the joint probability $p_{\text{data}}(\mathbf{x}, y)$ can be transformed into a deterministic classifier: $p_{\text{EBM}}(y|\mathbf{x}) \propto \exp(-U_\theta(\mathbf{x}, y))$. As a baseline for comparison, we additionally train a classifier $p_{\text{CLF}}(y|\mathbf{x})$ with the same architecture by maximising the conditional likelihood: $\mathbb{E}_{p_{\text{data}}}[\log p_{\text{CLF}}(y|\mathbf{x})]$. Results on the adult dataset can be seen in Figure 4. We find that the EBM and the baseline exhibit comparable accuracy. However, the baseline model is less calibrated, generating over-confident predictions. In contrast, the EBM learned through ED achieves better calibration, as evidenced by lower expected calibration error (Guo et al., 2017). Further details and results are provided in Appendix D.2.

6.3 Discrete Image Modelling

In this experiment, we evaluate our methods on high-dimensional binary spaces. Following the settings in Grathwohl et al. (2021), we conduct experiments on various image datasets and compare against contrastive divergence using various sampling methods, namely vanilla Gibbs sampling, Gibbs-With-Gradient (Grathwohl et al., 2021, GWG), Generative-Flow-Network (Zhang et al., 2022a, GFN), and Discrete-Unadjusted-Langevin-Algorithm (Zhang et al., 2022b, DULA). The training details are provided in Appendix D.3. After training, annealed importance sampling (Neal, 2001) is employed to estimate the negative log-likelihood (NLL).

Table 2 displays the NLLs on the test dataset. It is evident that energy discrepancy achieves comparable performance to the baseline methods on the Omniglot dataset. Despite the performance gap compared to the contrastive divergence methods on the MNIST dataset, energy discrepancy stands out for its efficiency, requiring only M evaluations of the energy function in parallel (see

https://doi.org/10.52202/079017-2516

Table 7 for the comparison of running time complexity). This represents a significant computational reduction compared to contrastive divergence, which lacks the advantage of parallelisation and involves simulating multiple MCMC steps. Additionally, our methods show superiority over CD-1 by a substantial margin, as demonstrated in Table 8, affirming the effectiveness of our approach. For further insights, we provide visualisations of the generated samples in Figure 10.

7 Conclusion and Limitations

In this paper we extend the training of energy-based models with energy discrepancy to discrete and mixed state spaces in a systematic way. We show that the energy-based model can be learned jointly on continuous and discrete variables and how prior assumptions about the geometry of the underlying discrete space can be utilised in the construction of the loss. Our method achieves promising results on a wide range of discrete modelling applications at a significantly lower computational cost than MCMC-based approaches. To the best of our knowledge, our approach is also the first working training method for energy-based models on tabular data sets, unlocking a wide range of inference applications for tabular data sets beyond the scope of classical joint energy-based models.

Limitations: Similar to prior work on energy discrepancy in continuous spaces (Schröder et al., 2023), our training method is sensitive to the assumption that the data distribution is positive on the whole state space. While our method scales to high-dimensional datasets like binary image data, where the positiveness of the data distribution is assumed to be violated due to the manifold hypothesis, the large difference between intrinsic and ambient dimensionality poses challenges to our approach and may explain why energy discrepancy cannot match the performance of contrastive divergence with a large number of MCMC steps on binary image data.

Broader Impact: In principle, our method can be used for imputation and prediction in tabular data sets and can thus have discriminating or excluding effects if used irresponsibly.

Outlook: For future work, we are interested in extensions to highly structured types of data such as molecules, text, or data arising from networks. So far, our work only considers cyclical and ordinal structures on the discrete space, while incorporating more complex structures as prior information into the rate function may be beneficial. Furthermore, interesting downstream applications ranging from table imputation with confidence bounds, simulation-based inference involving discrete variables, or reweighting of language models with residual EBMs have been left unexplored in this work.

Author Contributions

TS and ZO conceived the project idea to use an ED loss for the training of EBMs on discrete and mixed data. TS devised the main conceptual ideas, developed the theory, conducted the proofs and implemented the ED loss. ZO contributed to the conceptual ideas and designed and carried out the experiments. ABD supervised the conceptualisation and execution of the research project and contributed proof ideas. ZO, YL, and ABD checked derivations and proofs. TS and ZO equally contributed to the writing under the supervision of YL and ABD.

Acknowledgements

TS was supported by the EPSRC-DTP scholarship partially funded by the Department of Mathematics, Imperial College London. ZO was supported by the Lee Family Scholarship. We thank the anonymous reviewer for their comments.

References

Anderson, W. J. *Continuous-time Markov chains: An applications-oriented approach*. Springer Science & Business Media, 2012.

Besag, J. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 24(3):179–195, 1975.

Campbell, A., Benton, J., De Bortoli, V., Rainforth, T., Deligiannidis, G., and Doucet, A. A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 35:28266–28279, 2022.

Campbell, A., Yim, J., Barzilay, R., Rainforth, T., and Jaakkola, T. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. *International Conference on Machine Learning*, 41, 2024.

Carreira-Perpinan, M. A. and Hinton, G. On contrastive divergence learning. In *International workshop on artificial intelligence and statistics*, pp. 33–40. PMLR, 2005.

Ceylan, C. and Gutmann, M. U. Conditional noise-contrastive estimation of unnormalised models. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*. PMLR, 2018.

Dai, H., Singh, R., Dai, B., Sutton, C., and Schuurmans, D. Learning discrete energy-based models via auxiliary-variable local exploration. *Advances in Neural Information Processing Systems*, 33: 10443–10455, 2020.

Diaconis, P. and Stroock, D. Geometric bounds for eigenvalues of markov chains. *The annals of applied probability*, pp. 36–61, 1991.

Domke, J. Learning graphical model parameters with approximate marginal inference. *IEEE transactions on pattern analysis and machine intelligence*, 35(10):2454–2467, 2013.

Du, Y., Li, S., Tenenbaum, J., and Mordatch, I. Improved contrastive divergence training of energy based models. In *International Conference on Machine Learning*, 2021.

Eikema, B., Kruszewski, G., Dance, C. R., Elsahar, H., and Dymetman, M. An approximate sampler for energy-based models with divergence diagnostics. *Transactions of Machine Learning Research*, 2022.

Emami, P., Perreault, A., Law, J., Biagioni, D., and John, P. S. Plug & play directed evolution of proteins with gradient-based discrete MCMC. *Machine Learning: Science and Technology*, 4(2), 2023.

Foster, A., Jankowiak, M., O'Meara, M., Teh, Y. W., and Rainforth, T. A unified stochastic gradient approach to designing bayesian-optimal experiments. In *International Conference on Artificial Intelligence and Statistics*, pp. 2959–2969. PMLR, 2020.

Gao, R., Song, Y., Poole, B., Wu, Y. N., and Kingma, D. P. Learning energy-based models by diffusion recovery likelihood. *International Conference on Learning Representations*, 2021.

Grathwohl, W., Wang, K.-C., Jacobsen, J.-H., Duvenaud, D., Norouzi, M., and Swersky, K. Your classifier is secretly an energy based model and you should treat it like one. *International Conference on Learning Representations*, 2020.

Grathwohl, W., Swersky, K., Hashemi, M., Duvenaud, D., and Maddison, C. J. Oops I took a gradient: Scalable sampling for discrete distributions. In *International Conference on Machine Learning*, volume 38, 2021.

Gray, F. Pulse code communication. *United States Patent Number 2632058*, 1953.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1), 2012.

 https://doi.org/10.52202/079017-2516

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 297–304. JMLR Workshop and Conference Proceedings, 2010.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

Hinton, G. E. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

Hoogeboom, E., Nielsen, D., Jaini, P., Forré, P., and Welling, M. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*, 34:12454–12465, 2021.

Hyvärinen, A. Some extensions of score matching. *Computational statistics & data analysis*, 51(5): 2499–2512, 2007.

Jerrum, M. and Sinclair, A. Polynomial-time approximation algorithms for the ising model. *SIAM Journal on computing*, 22(5):1087–1116, 1993.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Kotelnikov, A., Baranchuk, D., Rubachev, I., and Babenko, A. Tabddpm: Modelling tabular data with diffusion models. In *International Conference on Machine Learning*, pp. 17564–17579. PMLR, 2023.

Lazaro-Gredilla, M., Dedieu, A., and George, D. Perturb-and-max-product: Sampling and learning in discrete energy-based models. *Advances in Neural Information Processing Systems*, 34:928–940, 2021.

Li, Y., Vinyals, O., Dyer, C., Pascanu, R., and Battaglia, P. Learning deep generative models of graphs. In *International Conference on Machine Learning*, 2018.

Liu, J., Kumar, A., Ba, J., Kiros, J., and Swersky, K. Graph normalizing flows. *Advances in Neural Information Processing Systems*, 32, 2019.

Liu, M., Liu, H., and Ji, S. Gradient-guided importance sampling for learning binary energy-based models. In *International Conference on Learning Representations*, 2023.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Losonczi, L. Eigenvalues and eigenvectors of some tridiagonal matrices. *Acta Mathematica Hungarica*, 60(3):309–322, 1992.

Lou, A., Meng, C., and Ermon, S. Discrete diffusion language modeling by estimating the ratios of the data distribution. *International Conference on Machine Learning*, 41, 2024.

Luo, Y., Yan, K., and Ji, S. Graphdf: A discrete flow model for molecular graph generation. In *International Conference on Machine Learning*, pp. 7192–7203. PMLR, 2021.

Lyu, S. Interpretation and generalization of score matching. In *Uncertainty in Artificial Intelligence*, volume 25, 2009.

Lyu, S. Unifying non-maximum likelihood learning objectives with minimum kl contraction. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Meng, C., Choi, K., Song, J., and Ermon, S. Concrete score matching: Generalized score matching for discrete data. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., and Dokania, P. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*, 33:15288–15299, 2020.

Neal, R. M. Annealed importance sampling. *Statistics and computing*, 11, 2001.

Nijkamp, E., Hill, M., Zhu, S.-C., and Wu, Y. N. Learning non-convergent non-persistent short-run MCMC toward energy-based model. In *Neural Information Processing Systems*, volume 33, 2019.

Niu, C., Song, Y., Song, J., Zhao, S., Grover, A., and Ermon, S. Permutation invariant graph generation via score-based generative modeling. In *International Conference on Artificial Intelligence and Statistics*, pp. 4474–4484. PMLR, 2020.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ou, Z., Xu, T., Su, Q., Li, Y., Zhao, P., and Bian, Y. Learning neural set functions under the optimal subset oracle. *Advances in Neural Information Processing Systems*, 35:35021–35034, 2022.

Papandreou, G. and Yuille, A. L. Perturb-and-map random fields: Using discrete optimization to learn and sample from energy models. In *2011 International Conference on Computer Vision*, pp. 193–200. IEEE, 2011.

Raghunathan, T. E. Synthetic data. *Annual review of statistics and its application*, 8:129–140, 2021.

Raginsky, M. Strong data processing inequalities and ϕ-sobolev inequalities for discrete channels. *IEEE Transactions on Information Theory*, 62(6):3355–3389, 2016.

Ramachandran, P., Zoph, B., and Le, Q. V. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

Rhodes, B. and Gutmann, M. U. Enhanced gradient-based mcmc in discrete spaces. *Transactions on Machine Learning Research*, 2022.

Schröder, T., Ou, Z., Lim, J., Li, Y., Vollmer, S., and Duncan, A. Energy discrepancies: a score-independent loss for energy-based models. *Advances in Neural Information Processing Systems*, 36, 2023.

Sen, P., Namata, G., Bilgic, M., Getoor, L., Galligher, B., and Eliassi-Rad, T. Collective classification in network data. *AI magazine*, 29(3):93–93, 2008.

Shi, C., Xu, M., Zhu, Z., Zhang, W., Zhang, M., and Tang, J. Graphaf: a flow-based autoregressive model for molecular graph generation. *International Conference on Learning Representations*, 2020.

Simonovsky, M. and Komodakis, N. Graphvae: Towards generation of small graphs using variational autoencoders. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part I 27*. Springer, 2018.

Sun, H., Dai, H., and Schuurmans, D. Optimal scaling for locally balanced proposals in discrete spaces. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022a.

Sun, H., Dai, H., Xia, W., and Ramamurthy, A. Path auxiliary proposal for MCMC in discrete space. In *International Conference on Learning Representations*, 2022b.

Sun, H., Dai, H., Dai, B., Zhou, H., and Schuurmans, D. Discrete Langevin samplers via Wasserstein gradient flow. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2023a.

Sun, H., Yu, L., Dai, B., Schuurmans, D., and Dai, H. Score-based continuous-time discrete diffusion models. In *The Eleventh International Conference on Learning Representations*, 2023b.

 https://doi.org/10.52202/079017-2516

Tee, G. J. Eigenvectors of block circulant and alternating circulant matrices. *New Zealand Journal of Mathematics*, 36(8):195–211, 2007.

Tieleman, T. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th international conference on Machine learning*, 2008.

Tran, T., Phung, D., and Venkatesh, S. Mixed-variate restricted boltzmann machines. In *Asian conference on machine learning*, pp. 213–229. PMLR, 2011.

Xiao, Z., Kreis, K., Kautz, J., and Vahdat, A. Vaebm: A symbiosis between variational autoencoders and energy-based models. *International Conference on Learning Representations*, 2020.

Xie, J., Lu, Y., Zhu, S.-C., and Wu, Y. A theory of generative convnet. In *International Conference on Machine Learning*, pp. 2635–2644. PMLR, 2016.

Xie, J., Lu, Y., Gao, R., and Wu, Y. N. Cooperative learning of energy-based model and latent variable model via MCMC teaching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Xie, J., Zhu, Y., Li, J., and Li, P. A tale of two flows: Cooperative learning of langevin flow and normalizing flow toward energy-based model. In *International Conference on Learning Representations*, 2022.

Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32, 2019.

You, J., Ying, R., Ren, X., Hamilton, W., and Leskovec, J. Graphrnn: Generating realistic graphs with deep auto-regressive models. In *International conference on machine learning*, pp. 5708–5717. PMLR, 2018.

Yueh, W.-C. Eigenvalues of several tridiagonal matrices. *Applied Mathematics E-Notes [electronic only]*, 5:66–74, 2005.

Zanella, G. Informed proposals for local MCMC in discrete spaces. *Journal of the American Statistical Association*, 115(530):852–865, 2020.

Zhang, D., Malkin, N., Liu, Z., Volokhova, A., Courville, A., and Bengio, Y. Generative flow networks for discrete probabilistic modeling. In *International Conference on Machine Learning*, pp. 26412–26428. PMLR, 2022a.

Zhang, P., Yin, H., Li, C., and Xie, X. Formulating discrete probability flow through optimal transport. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhang, R., Liu, X., and Liu, Q. A Langevin-like sampler for discrete distributions. In *International Conference on Machine Learning*, pp. 26375–26396. PMLR, 2022b.

Appendix for "Energy-Based Modelling for Discrete and Mixed Data via Heat Equations on Structured Spaces"

Contents

A Proofs of the Main Results

A.1 Proof of Theorem 1

Theorem 1. *Let $q_t(\cdot|x)$ be a Markov transition density defined by the rate matrix R with eigenvalues $0 = \lambda_1(R) \geq \lambda_2(R) \geq \cdots \geq \lambda_S(R)$ and uniform stationary distribution. Then, there exists a constant z_t* independent *of θ such that energy-discrepancy converges to the maximum-likelihood loss*

$$|\mathrm{ED}_{q_t}(p_{\mathrm{data}}, U_\theta) - \mathcal{L}_{\mathrm{MLE}}(\theta) - z_t| \leq \sqrt{S}\exp(-|\lambda_2(R)|t)\,\mathrm{KL}(p_{\mathrm{data}} \parallel p_\theta)\,.$$

with the loss of maximum-likelihood estimation $\mathcal{L}_{\mathrm{MLE}}(\theta) := -\mathbb{E}_{p_{\mathrm{data}}(\mathbf{x})}\big[\log p_\theta(\mathbf{x})\big]$.

Proof. For $\mathcal{X} = \{1, 2, \ldots, S\}$ and two probability distributions p_1 and p_2 on $\mathcal{X}$ define the marginal distributions

$$Q_t p_1(y) = \sum_{x\in\mathcal{X}} q_t(y|x)p_1(x)\,, \quad Q_t p_2(y) = \sum_{x\in\mathcal{X}} q_t(y|x)p_2(x)\,. \tag{10}$$

By the data-processing inequality it holds for all p_1, p_2 with $\mathrm{KL}(p_1 \parallel p_2) < \infty$ that

$$\mathrm{KL}(Q_t p_1 \parallel Q_t p_2) \leq c(t)\,\mathrm{KL}(p_1 \parallel p_2) \quad \text{with} \quad c(t) \leq 1\,. \tag{11}$$

We are going to bound the contraction rate $c(t)$ by firstly making use strong data processing inequality Raginsky (2016) which states that the contraction rate can be bounded by the Dobrushin contraction coefficient

$$c(t) \leq \theta(Q_t) = \sup_{x,x'} \|Q_t\delta_x - Q_t\delta'_x\|_{\mathrm{TV}}\,. \tag{12}$$

https://doi.org/10.52202/079017-2516

Furthermore, since total variation is a metric, the contraction between two points x, x' is upper bounded by the contraction towards the stationary distribution of the CTMC π:

$$\sup_{x,x'} \|Q_t\delta_x - Q_t\delta'_x\|_{\mathrm{TV}} \leq 2 \sup_x \|Q_t\delta_x - \pi\|_{\mathrm{TV}} \tag{13}$$

Next, we use Diaconis & Stroock (1991, Proposition 3) which states that for any $x \in \mathcal{X}$ and for eigenvalues $0 = \lambda_1(R) \geq \lambda_2(R) \geq \cdots \geq \lambda_S(R)$

$$\|Q_t\delta_{\mathbf{x}} - \pi\|^2_{\mathrm{TV}} \leq \frac{1}{4}\frac{1-\pi(\mathbf{x})}{\pi(x)} \exp(-2|\lambda_2(R)|t)\,. \tag{14}$$

The stationary distribution for $R^{\mathrm{ord}}, R^{\mathrm{cyc}}, R^{\mathrm{unif}}$ is given by the stationary distribution and hence $(1-\pi(\mathbf{x}))/\pi(x) \leq S$. Taking roots now yields

$$c(t) \leq \sup_{x,x'} \|Q_t\delta_x - Q_t\delta'_x\|_{\mathrm{TV}} \leq \sqrt{S} \exp(-|\lambda_2(R)|t)\,. \tag{15}$$

Finally, we conclude as in Schröder et al. (2023):

$$\begin{aligned}
\mathrm{KL}(Q_t p_{\mathrm{data}} \parallel Q_t p_\theta) &= \sum_{y\in\mathcal{X}} \left(\log Q_t p_{\mathrm{data}}(y) - \log \frac{Q_t p_\theta(y)}{p_\theta(x)} - \log p_\theta(x)\right) Q_t p_{\mathrm{data}}(y) \\
&= z_t + \sum_{y\in\mathcal{X}} U_{q_t,\theta}(y) Q_t p_{\mathrm{data}}(y) - U_\theta(x) - \log p_\theta(x)
\end{aligned}$$

with U independent entropy term $z_t := \sum_{y\in\mathcal{X}} Q_t p_{\mathrm{data}}(y) \log Q_t p_{\mathrm{data}}(y)$. After taking expectations with respect to $p_{\mathrm{data}}(x)$ we find

$$\begin{aligned}
0 &\leq \mathrm{KL}(Q_t p_{\mathrm{data}} \parallel Q_t p_\theta) \\
&= z_t + \sum_{y\in\mathcal{X}} U_{q_t,\theta}(y) Q_t p_{\mathrm{data}}(y) - \sum_{x\in\mathcal{X}} U_\theta(x) p_{\mathrm{data}}(x) - \sum_{x\in\mathcal{X}} \log p_\theta(x) p_{\mathrm{data}}(x) \\
&= z_t - \mathrm{ED}_{q_t}(p_{\mathrm{data}}, U_\theta) - \mathbb{E}_{p_{\mathrm{data}}(x)}[\log p_\theta(x)] \leq c(t)\,\mathrm{KL}(p_{\mathrm{data}} \parallel p_\theta)
\end{aligned}$$

□

A.2 Eigenvalue Decomposition of Rate Matrices for Proposition 1

The rate matrices for the cyclical and for the ordinal graph structures have a similar structure and are referred to as circulant and tridiagonal matrices. The easiest method for deriving the eigenvalue decompositions consists in deriving recurrence relations for the characteristic polynomial. A systematic study of block circulant matrices can be found in Tee (2007) and a study of tridiagonal matrices was given in Losonczi (1992); Yueh (2005). These more general results may be helpful when constructing perturbations for spaces with a more complex structure than the ones introduced in this work. We take already existing results and check that the desired results hold.

Proposition 1. *Assume the density $q_t(b|a) := q_t(y = b|x = a)$ is defined by the rate matrices R^{cyc} or R^{ord}. The transition density for all $a, b \in \{1, 2, \ldots, S\}$ is given by*

$$q_t^{\mathrm{cyc}}(b|a) = \frac{1}{S}\sum_{p=1}^{S} \exp(2\pi \mathrm{i} b\omega_p^{\mathrm{cyc}}) \exp\left((2\cos(2\pi\omega_p^{\mathrm{cyc}}) - 2)t\right) \exp(-2\pi \mathrm{i} a\omega_p^{\mathrm{cyc}}) \tag{8}$$

$$q_t^{\mathrm{ord}}(b|a) = \frac{2}{S}\sum_{p=1}^{S} \frac{1}{z_p} \cos((2b-1)\pi\omega_p^{\mathrm{ord}}) \exp\left((2\cos(2\pi\omega_p^{\mathrm{ord}}) - 2)t\right) \cos((2a-1)\pi\omega_p^{\mathrm{ord}}) \tag{9}$$

where $\omega_p^{\mathrm{cyc}} = (p-1)/S$ and $\omega_p^{\mathrm{ord}} = (p-1)/2S$, respectively, and $z_p = (2, 1, \ldots, 1)$.

Proof. In the circular case, the identity $R\mathbf{v}_p = \lambda_p \mathbf{v}_p$ reduces for a single row a to

$$v_{p,a-1} + v_{p,a+1} - 2v_{p,a} = \lambda_p v_{p,a} \tag{16}$$

For $v_{p,a} = \exp(-2\pi\mathrm{i}(p-1)a/S)/\sqrt{S}$ we find after dividing both sides by $v_{p,a}$:

$$\lambda_p = \exp\left(-2\pi\mathrm{i}\frac{(p-1)(a-1)-(p-1)a}{S}\right) + \exp\left(-2\pi\mathrm{i}\frac{(p-1)(a+1)-(p-1)a}{S}\right) - 2 \tag{17}$$
$$= 2\cos\left(2\pi\frac{p-1}{S}\right) - 2\,.$$

Furthermore, it is known that the Fourier basis is unitary, i.e. we have for $a, b \in \{1, 2, \ldots, S\}$

$$\sum_{p=1}^{S} \bar{v}_{a,p} v_{b,p} = \frac{1}{S}\sum_{p=1}^{S} \exp\left(-2\pi\mathrm{i}\frac{(a-b)(p-1)}{S}\right) = \delta(a,b)\,. \tag{18}$$

In the tridiagonal case, we have to study two cases. Firstly, we have to check in row 1 using $2\cos(x)\cos(y) = \cos(x+y) + \cos(x-y)$:

$$v_{p,2} - v_{p,1} = \lambda_p v_{p,1} = 2\cos\left(\frac{2\pi(p-1)}{2S}\right)\cos\left(\frac{\pi(p-1)}{2S}\right) - 2\cos\left(\frac{\pi(p-1)}{2S}\right) \tag{19}$$
$$= \cos\left(\frac{2\pi(p-1)}{2S} + \frac{\pi(p-1)}{2S}\right) + \cos\left(\frac{2\pi(p-1)}{2S} - \frac{\pi(p-1)}{2S}\right) - 2\cos\left(\frac{\pi(p-1)}{2S}\right)$$
$$= \cos\left(\frac{3\pi(p-1)}{2S}\right) - \cos\left(\frac{\pi(p-1)}{2S}\right)\,.$$

and equivalently for row S. In all other rows we have $v_{p,a+1} + v_{p,a-1} - 2v_{p,a} = \lambda_p v_{p,a}$ from

$$\frac{(2a-1)(p-1)}{2S} \pm \frac{2(p-1)}{2S} = \frac{(2(a\pm 1)-1)(p-1)}{2S}\,. \tag{20}$$

□

A.3 Proof of Scaling limit in Theorem 2

It is a typical generalisation of the central limit theorem that random walks attain Brownian motion as a universal scaling limit. We reproduce similar arguments for the law of the continuous time Markov chain. Without loss of generality we shift the state space by one and consider the state space $\{0, 1, \ldots, S-1\}$ with cyclical and ordinal structure and let $y_t \sim q_t(\cdot|x = \mu S)$, where we always assume that the process is initialised at state μS. Furthermore, we introduce the process z_t which is an unconstrained continuous time Markov chain on $\mathbb{Z}$ with rate matrix $R_{aa} = -2, R_{a,a+1} = 1, R_{a,a-1} = 1$. The constrained process can then be described in terms of the unconstrained one: Let $\varphi_S : \mathbb{Z} \to \{0, 1, \ldots, S-1\}$ with $\varphi_S(2nS + p) = p$ and $\varphi_S((2n+1)S + p) = -p$ for for $p \in \{0, 1, \ldots, S-1\}$ and $n \in \mathbb{Z}$. Then, φ_S reflects the unconstrained process z_t at the boundaries 0 and $S-1$, i.e. $y_t^{\mathrm{ord}} = \varphi_S(z_t)$ and $y_t^{\mathrm{cyc}} = z_t \bmod S$. For the unconstrained process we define the holding times, i.e. the random time intervals in which the process does not change state

$$h_a = \inf_{t\geq 0}\{t : z_t \neq x = a\} \tag{21}$$

and the jump process $N_t := \sum_{h\leq t} \delta(z_h \neq z_{h-})$, where $z_{h-} = \lim_{s\uparrow h} z_s$ which counts the number of state transitions up to time t. It is a standard result that the holding times are exponentially distributed (Anderson, 2012, Proposition 2.8) $h_a \sim \mathrm{Exp}(-R_{aa})$. Furthermore, since all holding times are identically distributed and $R_{aa} = -2$, the resulting jump process has Poisson distribution

$$N_t \sim \mathrm{Poisson}(2t)\,. \tag{22}$$

With these definitions, we can now first proof the Gaussian limit of z_t and then derive the limit of y_t.

Theorem 2 (Scaling limit)**.** *Let $y_t \sim q_t(\cdot|x = \mu S)$ with $\mu \in \{1/S, 2/S, \ldots, 1\}$, where q_t is either the transition density of the cyclical or ordinal perturbation. Let $\varphi : \mathbb{R} \to (0, 1]$, where for all $n \in \mathbb{Z}$ and $x \in (0, 1]$ $\varphi^{\mathrm{cyc}}(n + x) = x$ and $\varphi^{\mathrm{ord}}(2n + x) = x$, $\varphi^{\mathrm{ord}}(2n + 1 + x) = -x$. Then,*

$$y_{S^2 t}/S \xrightarrow{S\to\infty} \varphi(\xi_t) \quad \textit{with} \quad \xi_t \sim \mathcal{N}(\mu, 2t)\,.$$

https://doi.org/10.52202/079017-2516

Proof. First, we write z_t as the sum of independent increments starting at $x = \mu S$

$$z_t = \mu S + \sum_{j=1}^{N_t} J_j\,, \quad p(J_j = 1) = p(J_j = -1) = 1/2\,. \tag{23}$$

We can now compute the characteristic function $\chi_S(s) = \mathbb{E}[\exp(isz_t)]$ of z_t rescaled in space and time. We have the following:

$$\begin{aligned}
\chi_S(s) &= \mathbb{E}\left[\exp\left(\mathrm{i}s\frac{y_{S^2t}}{S}\right)\right] \\
&= \exp(\mathrm{i}s\mu)\mathbb{E}\left[\mathbb{E}\left[\exp\left(\mathrm{i}s\sum_{j=1}^{N_{S^2t}} J_j/S\right)\right]\middle| N_{S^2t}\right] \\
&= \exp(\mathrm{i}s\mu)\mathbb{E}\left[\prod_{j=1}^{N_{S^2t}} \cos\left(\frac{s}{S}\right)\middle| N_{S^2t}\right] \\
&= \exp(\mathrm{i}s\mu)\sum_{K=0}^{\infty}\cos\left(\frac{s}{S}\right)^K \frac{(2S^2t)^K\exp(-2S^2t)}{K!} = \exp(\mathrm{i}s\mu)\exp\left(2S^2t\cos\left(\frac{s}{S}\right) - 2S^2t\right)
\end{aligned} \tag{24}$$

where we used that $\cos(x) = 1/2(\exp(\mathrm{i}x) + \exp(-\mathrm{i}x))$ in the third step, the Poisson distribution of N_t in the fourth step, and the series expansion of the exponential function in the final step. Since $\cos(x) \approx 1 - 1/2x^2$, we now have point-wise for any s and due to the fact that $s/S \to 0$

$$\chi_S(s) \xrightarrow{S\to\infty} \exp(\mathrm{i}s\mu - ts^2) \tag{25}$$

which is the characteristic function of a Gaussian with variance $2t$ and mean μ. This proves the convergence in distribution of the rescaled unconstrained process z_t. Furthermore, for φ^{ord} and φ^{cyc} it holds $\varphi_S(z_t)/S = \varphi(z_t/S)$ for all $S \in \mathbb{N}$. Furthermore, $\varphi^{\mathrm{ord}}, \varphi^{\mathrm{cyc}}$ are continuous maps from $\mathbb{R}$ to $[0, 1)$ with reflecting or periodic boundary conditions. We thus have by the continuous mapping theorem

$$\frac{y_{S^2t}}{S} = \frac{\varphi_S(z_{S^2t})}{S} = \varphi\left(\frac{z_t}{S}\right) \xrightarrow{S\to\infty} \varphi(\xi_t) \tag{26}$$

with $\xi_t \sim \mathcal{N}(\mu, 2t)$. □

B Estimation of Energy Discrepancy

The energy discrepancy functional takes the form

$$\mathrm{ED}_q(p_{\mathrm{data}}, U) := \mathbb{E}_{p_{\mathrm{data}}(\mathbf{x})}[U(\mathbf{x})] - \mathbb{E}_{p_{\mathrm{data}}(\mathbf{x})}\mathbb{E}_{q(\mathbf{y}|\mathbf{x})}[U_q(\mathbf{y})]. \tag{27}$$

for a conditional perturbing density q and contrastive potential

$$U_q(\mathbf{y}) := -\log\sum_{\mathbf{x}'\in\mathcal{X}} q(\mathbf{y}|\mathbf{x}')\exp(-U(\mathbf{x}')) \tag{28}$$

While in theory energy discrepancy yields a valid training functional for energy-based models for almost any choice of conditional distribution q, the conditional distribution also needs to allow an estimation of U_q with low variance. This is particularly easy when q is symmetric, i.e. $q(\mathbf{y}|\mathbf{x}) = q(\mathbf{x}|\mathbf{y})$ in which case the contrastive potential can be expressed as an expectation which can readily be approximated from samples. This leads to

$$\mathcal{L}_{q,M,w}(U) := \frac{1}{N}\sum_{i=1}^{N}\log\left(w + \sum_{j=1}^{M}\exp(U(\mathbf{x}^i) - U(\mathbf{x}_-^{i,j}))\right) - \log(M) \tag{29}$$

with $\mathbf{x}^i \sim p_{\mathrm{data}}$, $\mathbf{y}^i \sim q_t(\cdot|\mathbf{x}^i)$, and $\mathbf{x}_-^{i,j} \sim q_t(\cdot|\mathbf{y}^i)$, where the offset w stabilises the loss approximation as discussed in Schröder et al. (2023). The interpretation of w is that contributions from negative samples with $U(\mathbf{x}_-^{i,j}) > U(\mathbf{x}^i)$ are exponentially suppressed as contributions to the loss functional, thus avoiding the energies of negative samples to explode.

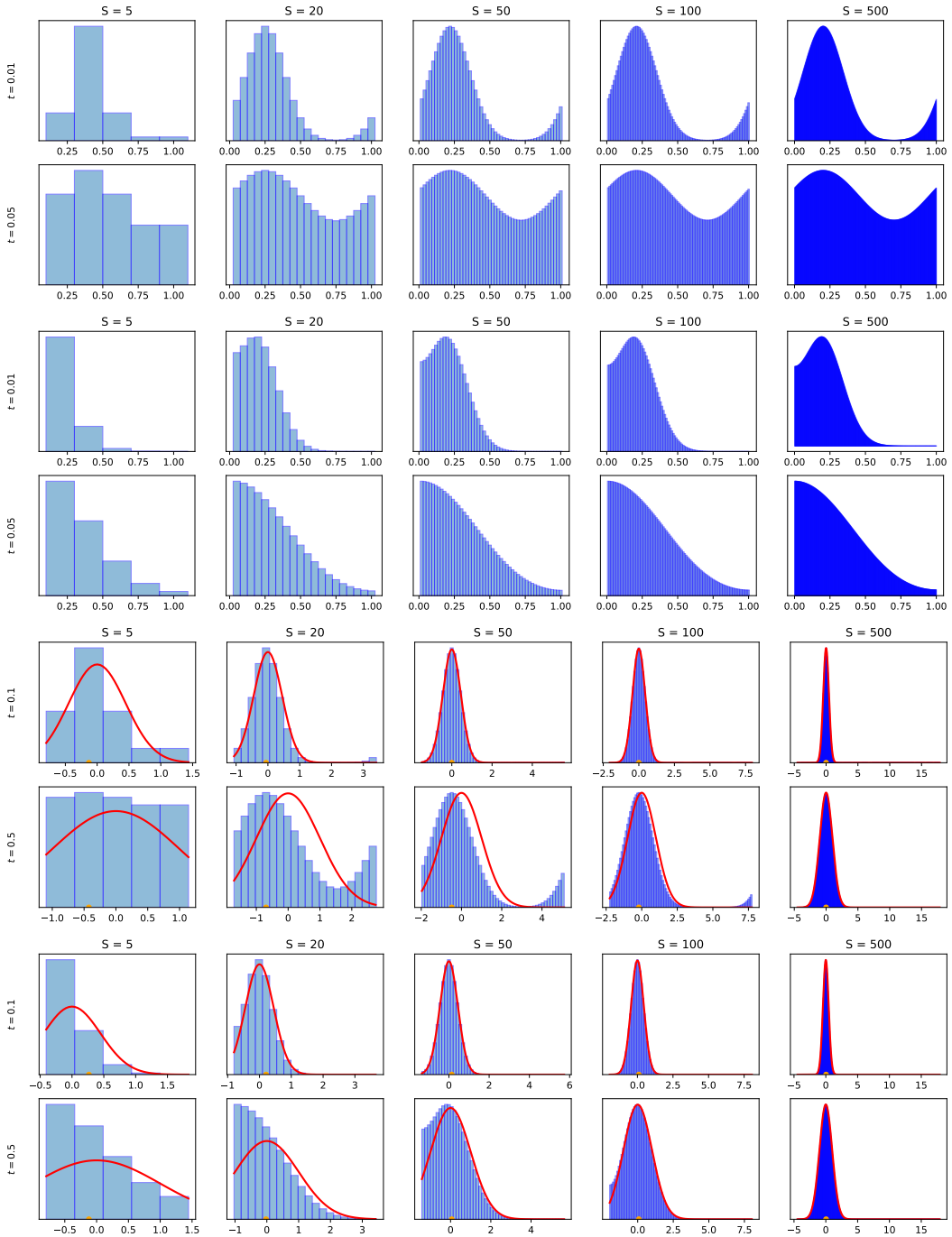


Figure 5: Scaling limit of the introduced perturbations. Top: Convergence of rescaled cyclical and ordinal perturbations y_{S^2t}/S for base time parameters $t = 0.01$ and $t = 0.05$ to Gaussian $[0, 1)$ with non-trivial boundary conditions. One can see that the perturbation converges to a fixed shape on the normalised state space. Bottom: Convergence of rescaled cyclical and ordinal perturbation $(y_{St} - \mathbb{E}[y_{St}])/\sqrt{S}$ for base time parameters $t = 0.1$ and $t = 0.5$ to Gaussian on $\mathbb{R}$ (red line). The orange mark indicates the initial state. One can see that the perturbation remains non-trivial as the state space grows to infinity at rate $\sqrt{S}$.

https://doi.org/10.52202/079017-2516

B.1 Binary Case

On binary spaces, the construction of perturbations is particularly simple. We give some details in this subsection.

Bernoulli perturbation. As proposed previously in (Schröder et al., 2023, Appendix B.3), q can be defined as a Bernoulli distribution. Specifically, for $\boldsymbol{\xi} \sim \text{Bernoulli}(\varepsilon)^d, \varepsilon \in (0,1)$ define the Bernoulli perturbed data point as $\mathbf{y} = \mathbf{x} + \boldsymbol{\xi} \mod 2$. This induces a symmetric transition density $q(\mathbf{y}|\mathbf{x})$ on $\{0,1\}^d$. The Bernoulli random variable $\boldsymbol{\xi}_k$ emulates an indicator function, signifying in each dimension whether to flip the entry of $\mathbf{x}$. The value of ϵ controls the information loss induced by the perturbation. In theory, larger values of ϵ lead to a more data-efficient loss, while smaller values of ϵ may be more practical as they contribute to improved training stability.

It is easy to compute the matrix exponential

$$\exp\left(t\begin{pmatrix}-1 & 1\\ 1 & -1\end{pmatrix}\right) = \frac{1}{2}\begin{pmatrix}1+e^{-2t} & 1-e^{-2t}\\ 1-e^{-2t} & 1+e^{-2t}\end{pmatrix} \xrightarrow{t\to\infty} \begin{pmatrix}1/2 & 1/2\\ 1/2 & 1/2\end{pmatrix} \tag{30}$$

thus relating the continuous time Markov chain framework on $\{0,1\}$ to a Bernoulli perturbation with parameter $0.5 * (1 - e^{-2t})$.

Neighbourhood-based perturbation and grid perturbation. Inspired by concrete score matching (Meng et al., 2022), one can introduce a perturbation scheme based on neighbourhood maps: $\mathbf{x} \mapsto \mathcal{N}(\mathbf{x})$, which assigns each data point $\mathbf{x} \in \mathcal{X}$ a set of neighbours $\mathcal{N}(\mathbf{x})$. In this case, the forward transition density is given by the uniform distribution over the set of neighbours, *i.e.,* $q(\mathbf{y}|\mathbf{x}) = \frac{1}{|\mathcal{N}(\mathbf{x})|}\delta_{\mathcal{N}(\mathbf{x})}(\mathbf{y})$. A special case is the grid neighbourhood

$$\mathcal{N}_{\text{grid}}(\mathbf{x}) = \{\mathbf{y} \in \{0,1\}^d \,:\, \mathbf{y} - \mathbf{x} = \pm\mathbf{e}_k,\, k = 1,2,\ldots,d\}, \tag{31}$$

where $\mathbf{e}_k$ is a vector of zeros with a one in the k-th entry. Notably, this neighbourhood structure also exhibits symmetry, *i.e.,* $\mathcal{N}^{-1}_{\text{grid}}(\mathbf{x}) = \mathcal{N}_{\text{grid}}(\mathbf{x})$. The same perturbation can be derived from an Euler discretisation of the continuous time Markov chain. On a binary space we have for $t = 1$ for the same rate matrix as in Equation (30)

$$F := \exp(R) \approx \text{id} + \begin{pmatrix}-1 & 1\\ 1 & -1\end{pmatrix} = \begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix} \tag{32}$$

which is a completely deterministic perturbation which always changes the state of the data point. Combining this with the localisation to a grid we find

$$q(\mathbf{y}|\mathbf{x}) = \sum_{k=1}^{d}\frac{1}{d}F_{yx} = \begin{cases}1/d & \mathbf{y}-\mathbf{x} = \pm\mathbf{e}_k\\ 0 & \text{else}\end{cases} \tag{33}$$

thus recovering the grid perturbation.

B.2 Connection to pseudo-likelihood estimation

Define $q(\mathbf{y}|\mathbf{x}) = \frac{1}{d}\sum_{k=1}^{d}\delta(y_k, \square)\delta(\mathbf{y}_{\neg k}, \mathbf{x}_{\neg k})$ which masks exactly one entry of the data vector. For simiplicity, we write $\mathbf{y} = \mathbf{x}_{\neg k}$ for the masked state and ommit the $\square$. Then, energy discrepancy is given for a sampled perturbation $\mathbf{y} = \mathbf{x}_{\neg k}$ as

$$\begin{aligned} &U_\theta(\mathbf{x}) + \log\sum_{\mathbf{x}\in\mathcal{X}} q(\mathbf{y} = \mathbf{x}_{\neg k}|\mathbf{x})\exp(-U_\theta(\mathbf{x})) \\ &= -\log\frac{\exp(-U_\theta(\mathbf{x}))}{\sum_{s\in\{1,2,\ldots,S_k\}}\exp(-U_\theta(x_1,\ldots,x_k=s,\ldots,x_d))} = -\log p_\theta(x_k|\mathbf{x}_{\neg k}) \end{aligned} \tag{34}$$

Hence, this specific ED loss function is indeed a Monte Carlo approximation of pseudo-likelihood. Energy discrepancy offers additional flexibility through the tunable choice of t and M, thus making ED adaptable to the structure of the underlying space and more efficient in practice, since the normalisation of the pseudo-likelihood is only computed from M samples and does not require the integration along an entire state space dimension.

Algorithm 1 Training

perturb the training sample $\mathbf{x}$ using (35)
$\mathbf{y} \sim q_t(\cdot|\mathbf{x})$
generate M negative samples
$\mathbf{x}_-^i \sim q_t(\cdot|\mathbf{y}), \quad i = 1, 2, \ldots, M$
evaluate the energy difference
$d_\theta \leftarrow \frac{1}{M} \sum_{i=1}^{M} \exp(U_\theta(\mathbf{x}) - U_\theta(\mathbf{x}_-^i))$
update parameter θ using (29)
$\theta \leftarrow \theta - \eta \nabla_\theta \log(w/M + d_\theta)$

Algorithm 2 Sampling

initialise samples
$\mathbf{x}^{\text{num}} \sim \mathcal{N}(0, \mathbf{I})$; $\mathbf{x}^{\text{cat}} \sim \bigotimes_{k=1}^{d_{\text{cat}}} \text{Uniform}(S_k)$
for $1, \ldots, N$ **do**
  for $k \leftarrow 1, \ldots, d_{\text{cat}}$ **do**
    update numerical features using (36)
    $\mathbf{x}^{\text{num}} \leftarrow \mathbf{x}^{\text{num}} - \frac{\epsilon}{2} U_\theta(\mathbf{x}) + \sqrt{\epsilon}\boldsymbol{\omega}$, $\boldsymbol{\omega} \sim \mathcal{N}(0, \mathbf{I})$
    update categorical features using (37)
    $x_k^{\text{cat}} \leftarrow x_k^{\text{cat}} \sim p_\theta(x_k^{\text{cat}} | \mathbf{x}^{\text{num}}, \mathbf{x}_{\neg k}^{\text{cat}})$
  end for
end for

Figure 6: The training and sampling procedures of energy discrepancy on tabular data. We use one training sample only to illustrate.

C Tabular Data Synthesising with Energy-Based Models

In this section, we introduce how to use energy discrepancy for training an energy-based model on tabular data. Let d_{num} and d_{cat} be the number of numerical columns and categorical columns, respectively. Each row in the table is a data point represented as a vector of numerical features and categorical features $\mathbf{x} = [\mathbf{x}^{\text{num}}, \mathbf{x}^{\text{cat}}]$, where $\mathbf{x}^{\text{num}} \in \mathbb{R}^{d_{\text{num}}}$ and $\mathbf{x}^{\text{cat}} \in \bigotimes_{k=1}^{d_{\text{cat}}} \{1, \ldots, S_k\}$. To train an EBM with energy discrepancy, one should define the perturbation methods, which can be done by solving the differential equation in (6). For the numerical features, we choose the Gaussian perturbation as in Schröder et al. (2023), which has the transition probability in the form of

$$q_t^{\text{num}}(\mathbf{y}^{\text{num}}|\mathbf{x}^{\text{num}}) = \mathcal{N}(\mathbf{y}^{\text{num}}|\mathbf{x}^{\text{num}}, t\mathbf{I}).$$

For the categorical features, we perturb each attribution independently:

$$q_t^{\text{cat}}(\mathbf{y}^{\text{cat}}|\mathbf{x}^{\text{cat}}) = \prod_{k=1}^{d_{\text{cat}}} q_t(y_k^{\text{cat}}|x_k^{\text{cat}}).$$

Accordingly, there are three different perturbation methods for $q_t(y_k^{\text{cat}}|x_k^{\text{cat}})$:

- Uniform perturbation defined in (7):

$$q_t^{\text{uni}}(y_k^{\text{cat}}|x_k^{\text{cat}}) = e^{-t}\delta(y_k^{\text{cat}}, x_k^{\text{cat}}) + \frac{1 - e^{-t}}{S} \sum_{a=1}^{S} \delta(y_k^{\text{cat}}, a)$$

  Here, the time parameter was rescaled to be independent of S for ease of implementation.

- Cyclical perturbation defined in (8):

$$q_t^{\text{cyc}}(y_k^{\text{cat}}|x_k^{\text{cat}}) = \frac{1}{S} \sum_{p=1}^{S} \exp(2\pi \mathrm{i} y_k^{\text{cat}} \omega_p^{\text{cyc}}) \exp\left((2\cos(2\pi\omega_p^{\text{cyc}}) - 2)t\right) \exp(-2\pi \mathrm{i} x_k^{\text{cat}} \omega_p^{\text{cyc}})$$

- Ordinal perturbation defined in (9):

$$q_t^{\text{ord}}(y_k^{\text{cat}}|x_k^{\text{cat}}) = \frac{2}{S} \sum_{p=1}^{S} \frac{1}{z_p} \cos((2y_k^{\text{cat}} - 1)\pi\omega_p^{\text{ord}}) \exp\big((2\cos(2\pi\omega_p^{\text{ord}}) - 2)t\big) \cos((2x_k^{\text{cat}} - 1)\pi\omega_p^{\text{ord}})$$

To reduce the scale of noise, we further introduce grid perturbation, which involves perturbing only one attribute at a time

$$q_t^{\text{cat}}(\mathbf{y}^{\text{cat}}|\mathbf{x}^{\text{cat}}) = \frac{1}{d_{\text{cat}}} \sum_{k=1}^{d_{\text{cat}}} q_t(y_k^{\text{cat}}|x_k^{\text{cat}}).$$

Theoretically, grid perturbation can be used alongside any type of perturbation described in (7, 8, 9). In our experimental studies, we only explore the combination of grid perturbation with uniform

https://doi.org/10.52202/079017-2516

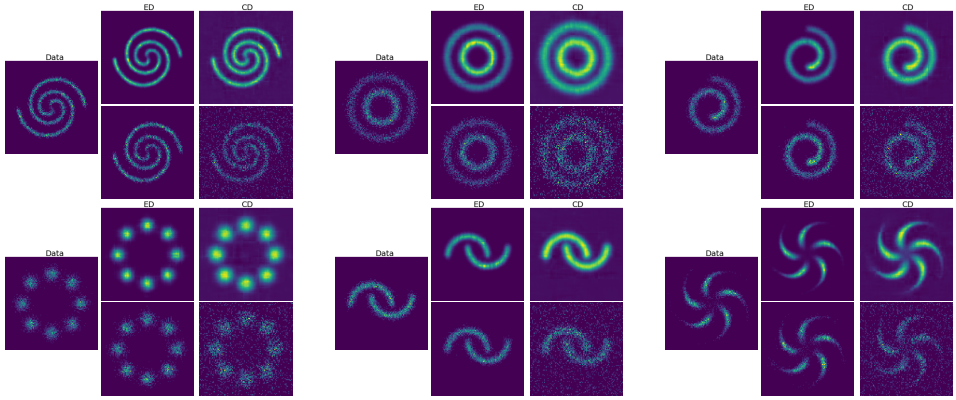


Figure 7: Additional density estimation results on the dataset with 16 dimensions and 5 states.

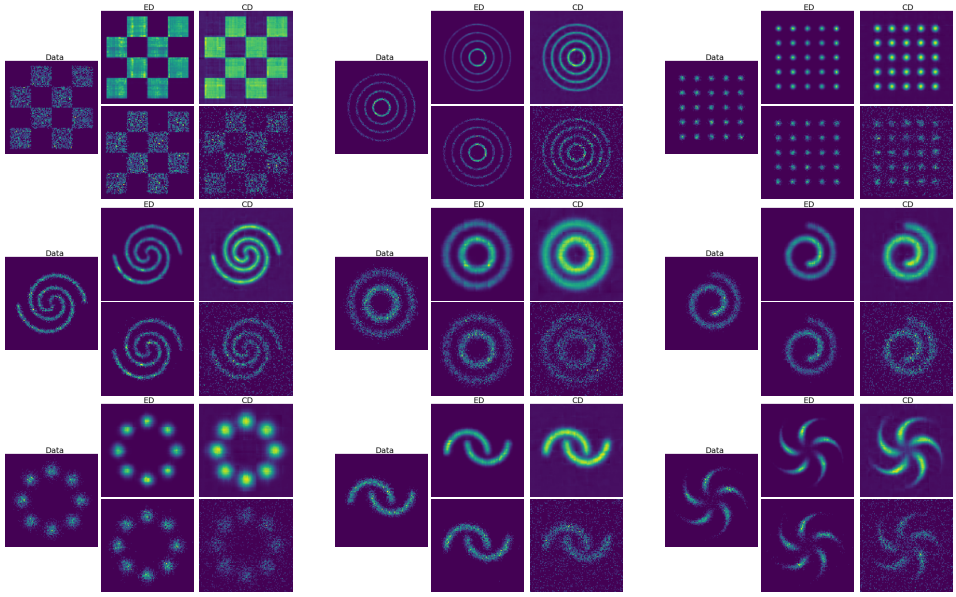


Figure 8: Additional density estimation results on the dataset with 12 dimensions and 10 sates.

perturbation. By combining the Gaussian perturbation and categorical perturbation together, we can then draw the negative samples $\mathbf{x}_-$ via $\mathbf{y} \sim q_t(\cdot|\mathbf{x})$, and $\mathbf{x}_- \sim q_t(\cdot|\mathbf{y})$, where

$$q_t(\mathbf{y}|\mathbf{x}) = q^{\text{num}}_{t_{\text{num}}}(\mathbf{y}^{\text{num}}|\mathbf{x}^{\text{num}}) q^{\text{cat}}_{t_{\text{cat}}}(\mathbf{y}^{\text{cat}}|\mathbf{x}^{\text{cat}}). \tag{35}$$

Therefore, the energy function U_θ can be learned by minimising the loss function in (29). We summarise the training procedure in Algorithm 1.

After training, new tabular data is synthesised by alternately applying Langevin dynamics and Gibbs sampling. Specifically, we update the numerical feature $\mathbf{x}^{\text{num}}$ via Langevin dynamics

$$\mathbf{x}^{\text{num}} \leftarrow \mathbf{x}^{\text{num}} - \frac{\epsilon}{2} \nabla_{\mathbf{x}^{\text{num}}} U_\theta([\mathbf{x}^{\text{num}}, \mathbf{x}^{\text{cat}}]) + \sqrt{\epsilon}\boldsymbol{\omega}, \quad \boldsymbol{\omega} \sim \mathcal{N}(0, \mathbf{I}) \tag{36}$$

For the categorical feature $\mathbf{x}^{\text{cat}}$, we employ Gibbs sampler

$$x^{\text{cat}}_k \sim p_\theta(x^{\text{cat}}_k|\mathbf{x}^{\text{num}}, \mathbf{x}^{\text{cat}}_{\neg k}) \propto \exp(-U_\theta([\mathbf{x}^{\text{num}}, x^{\text{cat}}_k, \mathbf{x}^{\text{cat}}_{\neg k}])), \quad k = 1, 2, \ldots, d_{\text{cat}}, \tag{37}$$

where $\mathbf{x}^{\text{cat}}_{\neg k}$ denotes the vector $\mathbf{x}^{\text{cat}}$ excluding the k-th attribute. The sampling procedure is summarised in Algorithm 2.

Table 3: Experimental results of discrete density estimation. We display the negative log-likelihood (NLL). The results of baselines are taken from Zhang et al. (2022a).

Metric	Method	2spirals	8gaussians	circles	moons	pinwheel	swissroll	checkerboard
NLL↓	PCD	20.094	19.991	20.565	19.763	19.593	20.172	21.214
	ALOE+	20.062	19.984	20.570	19.743	19.576	20.170	21.142
	EB-GFN	20.050	19.982	**20.546**	19.732	**19.554**	20.146	20.696
	ED-Bern	**20.039**	19.992	20.601	**19.710**	19.568	**20.084**	20.679
	ED-Grid	20.049	**19.965**	20.601	19.715	19.564	20.088	**20.678**

Table 4: Experimental results of discrete density estimation. We display the MMD (in units of 1×10^{-4}). The results of baselines are taken from Zhang et al. (2022a).

Metric	Method	2spirals	8gaussians	circles	moons	pinwheel	swissroll	checkerboard
MMD↓	PCD	2.160	0.954	0.188	0.962	0.505	1.382	2.831
	ALOE+	0.149	0.078	0.636	0.516	1.746	0.718	12.138
	EB-GFN	0.583	0.531	0.305	0.121	0.492	0.274	**1.206**
	ED-Bern	0.120	0.014	0.137	**-0.088**	**0.046**	**0.045**	1.541
	ED-Grid	**0.097**	**-0.066**	**0.022**	0.018	0.351	0.097	2.049

D Additional Experimental Results

In this section, we present the detailed experimental settings and additional results. All experiments are conducted on a single Nvidia RTX 3090 GPU with 24GB of VRAM.

D.1 Discrete Density Estimation

Experimental Details. This experiment keeps a consistent setting with Dai et al. (2020). We first generate 2D floating-points from a continuous distribution $\hat{p}$ which lacks a closed form but can be easily sampled. Then, each sample $\hat{\mathbf{x}} := [\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2] \in \mathbb{R}^2$ is converted to a discrete data point $\mathbf{x} \in \{0,1\}^{32}$ using Gray code. To be specific, given $\hat{\mathbf{x}} \sim \hat{p}$, we quantise both $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$ into 16-bits binary representations via Gray code (Gray, 1953), and concatenate them together to obtain a 32-bits vector $\mathbf{x}$. As a result, the probabilistic mass function in the discrete space is $p(\mathbf{x}) \propto \hat{p}\left([\mathrm{GrayToFloat}(\mathbf{x}_{1:16}), \mathrm{GrayToFloat}(\mathbf{x}_{17:32})]\right)$. To extend datasets beyond binary cases, we adhere to the same protocol but utilise base transformation instead. This transformation enables the conversion of floating-point coordinates into discrete variables with different state sizes. It is important to highlight that learning EBMs in such discrete spaces presents challenges due to the highly non-linear characteristics of both the Gray code and base transformation.

We parameterise the energy function using a 4 layer MLP with 256 hidden dimensions and Swish (Ramachandran et al., 2017) activation. To train the EBM, we adopt the Adam optimiser with a learning rate of 0.0001 and a batch size of 128 to update the parameter for 10^5 steps. For the energy discrepancy, we choose $w = 1, M = 32$ and the grid perturbation for all variants. For contrastive divergence, we employ short-run MCMC using Gibbs sampling with 10 rounds (i.e., $10 * S$ steps).

After training, we quantitatively evaluate all methods using the negative log-likelihood (NLL) and the maximum mean discrepancy (MMD). To be specific, the NLL metric is computed based on $4,000$ samples drawn from the data distribution, and the normalisation constant is estimated using importance sampling with $1,000,000$ samples drawn from a variational Bernoulli distribution with $p = 0.5$. For the MMD metric, we follow the setting in Zhang et al. (2022a), which adopts the exponential Hamming kernel with 0.1 bandwidth. Moreover, the reported performances are averaged over 10 repeated estimations, each with $4,000$ samples, which are drawn from the learned energy function via Gibbs sampling.

Qualitative Results. In Figures 7 and 8, we present additional qualitative results of the learned energy on datasets with 5 and 10 states. We see that ED consistently yields more accurate energy landscapes compared to CD. Notably, we only showcase results using grid perturbation with the uniform rate matrix, as qualitative findings are consistent across different perturbation methods. Additionally, we

https://doi.org/10.52202/079017-2516

Table 5: Statistics of the real-world datasets.

Dataset	# Rows	# Num	# Cat	# Train	# Validation	# Test	Task Type
Adult	48,842	6	9	28,943	3,618	16,281	Binary Classification
Bank	45,211	7	10	36,168	4,521	4,522	Binary Classification
Cardio	70,000	5	6	56,000	7,000	7,000	Binary Classification
Churn	7,043	3	15	5,634	704	705	Binary Classification
Mushroom	61,096	3	17	48,855	6,107	6,107	Binary Classification
Beijing	43,824	7	5	35,058	4,383	4,383	Regression

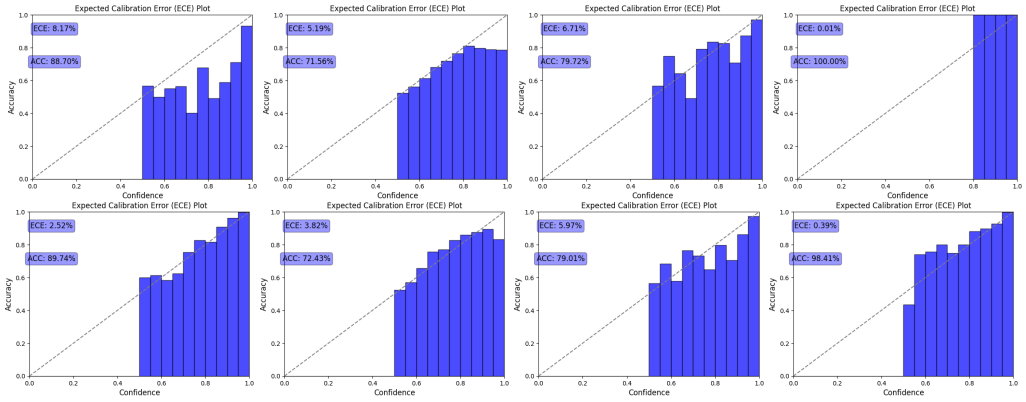


Figure 9: Comparison of calibration results between the baseline (top) and energy discrepancy (bottom) on varying datasets. Left to right: Bank, Cardio, Churn, Mushroom.

empirically observe that gradient-based Gibbs sampling methods (Grathwohl et al., 2021; Zhang et al., 2022b) tend to generate samples outside the data support more readily. In this regard, we only display the results of CD methods using vanilla Gibbs sampling.

Quantitative Results. The quantitative results are illustrated in Tables 3 and 4, indicating the superior performance of our approaches in most scenarios. Notably, previous studies on discrete EBM modelling exclusively focus on binary cases. As a result, we only present the quantitative comparison for the dataset with 2 states.

D.2 Tabular Data Synthesising

Experimental Details for the Synthetic Dataset. For the synthetic dataset, we parametrise the energy function using three MLP layers with 256 hidden states and Swish activation. To handle mixed data types, we transform each categorical feature into a 4-dimensional embedding using a linear layer, and then concatenate these embeddings with the numerical features as input. To train the EBM, we apply the Adam optimiser with a learning rate of 0.0001 and a batch size of 128. We update the parameters over $1,000$ epochs, with each epoch consisting of 100 update iterations. For ED, we set $w = 1, M = 32$, using Gaussian perturbation for the numerical features and grid perturbation for the categorical features. For CD, we incorporate the replay buffer strategy and employ Langevin dynamics and Gibbs sampling with 50 rounds (totalling $50 * S$ steps) for the numerical and categorical features, respectively.

Experimental Details for the Real-world Dataset. Table 5 summarises the statistical properties of the datasets. To parameterise the energy function and handle mixed data types, we use the same approach but with 1024 hidden units instead of 256. We train the model using the AdamW optimiser (Loshchilov & Hutter, 2019) with a learning rate of 0.0001 and a weight decay rate of 0.0005. The model is trained for $20,000$ update steps with a batch size of 4096. For ED, Gaussian perturbation is employed for numerical features, while categorical features undergo different perturbation methods. Specifically, TabED-Uni and TabED-Grid use uniform and grid perturbations with $t = 0.1$, respectively. For TabED-Cyc and TabED-Ord, corresponding to cyclical and ordinal perturbations, quadratic scaling is applied with t chosen from the best performance in $\{0.01, 0.005, 0.001\}$. Moreover, CD

Table 6: Results on density similarity between the synthesis and real tabular data.

Methods	Single-column Density Similarity ↑						Pair-wise Correlation Similarity ↑					
	Adult	Bank	Cardio	Churn	Mushroom	Beijing	Adult	Bank	Cardio	Churn	Mushroom	Beijing
CTGAN	.814	.866	.906	.901	.951	.799	.744	.769	.717	.826	.828	.761
TVAE	.783	.824	.892	.899	.965	.711	.669	.772	.687	.808	.919	.618
TabCD	.719	.790	.824	.845	.618	.799	.522	.600	.629	.710	.428	.761
TabDDPM	.988	.998	.992	.976	.987	.980	.975	.894	.870	.953	.962	.946
TabED-Uni	.785	.779	.914	.886	.878	.933	.653	.683	.783	.808	.770	.793
TabED-Grid	.751	.766	.945	.846	.951	.951	.583	.768	.829	.764	.842	.842
TabED-Cyc	.778	.826	.937	.834	.969	.751	.636	.703	.810	.755	.860	.685
TabED-Ord	.828	.894	.933	.887	.943	.747	.702	.796	.811	.791	.826	.662
TabED-Str	.77	.798	–	–	–	.892	.632	.660	–	–	–	.759

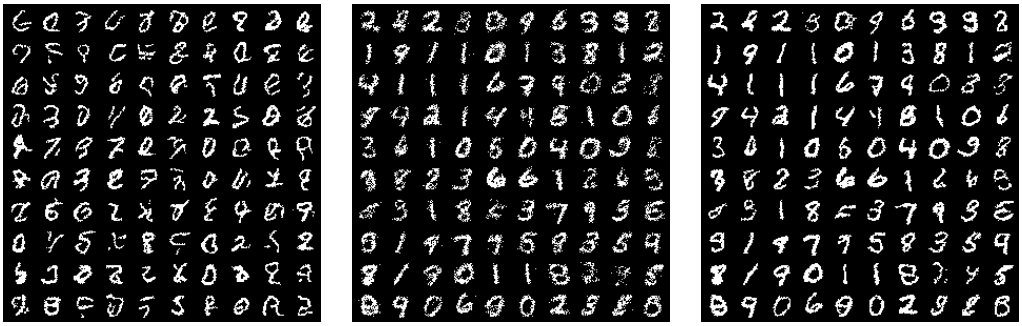


Figure 10: Dynamic MNIST samples generated from the learned energy function using ED-Grid with various sampling methods. Left to right: GWG, GFlowNet, GFlowNet+GWG.

utilises the same algorithm as in the synthetic dataset, but with 10 steps for short-run MCMC. The reported results are averaged over 10 randomly sampled synthetic data.

Experimental Details for Calibration. Let y and $\mathbf{x}$ be the target label and the rest features in the tabular data, we can transform a learned EBM $U_\theta(\mathbf{x}, y)$ into a deterministic classifier: $p_{\text{EBM}}(y|\mathbf{x}) \propto \exp(-U_\theta(\mathbf{x}, y))$. As a baseline for comparison, we additionally train a classifier $p_{\text{CLF}}(y|\mathbf{x})$ with the same architecture by maximising the conditional likelihood: $\mathbb{E}_{p_{\text{data}}}[\log p_{\text{CLF}}(y|\mathbf{x})]$. In particular, we utilise the Adam optimiser with a learning rate of 0.001 and a batch size of 4096 to train the classifier p_{CLF}. The model undergoes training for 50 epochs.

Additional Results for Calibration. Figure 9 presents additional calibration results across different datasets. It shows that the energy-based classifier learned by energy discrepancy exhibits superior calibration compared to the deterministic classifier, except for the Mushroom dataset, where the deterministic classifier achieves 100% accuracy, resulting in low calibration error.

Additional Results with Other Metrics. We evaluate our methods against baselines using two additional metrics: single-column density similarity and pair-wise correlation similarity. These metrics assess the similarity in the empirical distribution of individual columns and the correlations between pairs of columns in the generated versus real tables. Both metrics can be computed using the open-source SDMetrics API. As shown in Table 6, the result shows that the proposed ED-based approaches either outperform or achieve comparable performance to the baselines across most datasets.

D.3 Discrete Image Modelling

Experimental Details. In this experiment, we parametrise the energy function using ResNet (He et al., 2016) following the settings in Grathwohl et al. (2021); Zhang et al. (2022b), where the network has 8 residual blocks with 64 feature maps. Each residual block has 2 convolutional layers and uses Swish activation function (Ramachandran et al., 2017). We choose $M = 32, w = 1$ for all variants of energy discrepancy, $\epsilon = 0.001$ in Bernoulli perturbations. Note that here we choose a relatively small ϵ since we empirically find that the loss of energy discrepancy converges to a constant rapidly with

https://doi.org/10.52202/079017-2516

Table 7: Running time complexity comparison for energy discrepancy and contrastive divergence.

Time \ Method	CD-1	CD-5	CD-10	ED-Bern	ED-Grid
Per Iteration (s)	0.0583	0.1904	0.3351	0.0905	0.0872
Per Epoch (s)	29.1660	95.2178	167.5718	46.4317	44.0621

Table 8: Experimental results of the comparison between energy discrepancy and contrastive divergence with varying MCMC steps.

Dataset \ Method	CD-1	CD-3	CD-5	CD-7	CD-10	ED-Bern	ED-Grid
Static MNIST	182.53	130.94	102.70	98.07	**88.13**	96.11	90.61
Dynamic MNIST	157.14	130.56	97.50	91.00	**84.16**	97.12	90.19
Omniglot	nan.	161.96	142.91	149.68	146.11	97.57	**93.94**

larger ϵ, which can not provide meaningful gradient information to update the parameters. All models are trained with Adam optimiser with a learning rate of 0.0001 and a batch size of 100 for $50,000$ iterations. We perform model evaluation every $5,000$ iteration by conducting Annealed Importance Sampling (AIS) with the GWG (Grathwohl et al., 2021) sampler for $10,000$ steps. The reported results are obtained from the model that achieves the best performance on the validation set. After training, we finally report the negative log-likelihood by running $300,000$ iterations of AIS.

Qualitative Results. To qualitatively assess the validity of the learned EBM, this study presents generated samples from the dynamic MNIST dataset. We first train an EBM using ED-Grid and then synthesise samples by employing various sampling methods, including: i) GWG (Grathwohl et al., 2021) with 1000 steps; ii) GFlowNet with the same architecture and training procedure as per Zhang et al. (2022a); and iii) GFlowNet followed by GWG with 100 steps.

Empirically, we find that the quality of generated samples can be improved with more advanced sampling approaches. As depicted in Figure 10, the GWG sampler suffers from mode collapse, leading to samples with similar patterns. In other hands, GFlowNet enhances the quality to some extent, but it produces noisy images. To address this issue, we apply GWG with 100 steps following the GFlowNet. It can be seen that the resulting GFlowNet+GWG sampler yields the highest quality with clear digits. These observations validate the capability of our energy discrepancies to accurately learn the energy landscape from high-dimensional datasets. We leave the development of a more advanced sampler in future work to further improve the quality of generated images using our energy discrepancy approaches.

Time Complexity Comparison for Energy Discrepancy and Contrastive Divergence. Energy discrepancy offers greater training efficiency than contrastive divergence, as it does not rely on MCMC sampling. In this experiment, we evaluate the running time per iteration and epoch for energy discrepancy and contrastive divergence in training a discrete EBM on the static MNIST dataset. The experiments include contrastive divergence with varying MCMC steps and variants of energy discrepancy with a fixed value of $M = 32$. The results, presented in Table 7, highlight that ED-Bern and ED-Grid are the fastest options, as they do not involve gradient computations during training.

Comparison to Contrastive Divergence with Different MCMC Steps. Considering the greater training efficiency of energy discrepancy over contrastive divergence, this study comprehensively compares these two methods with varying MCMC steps in contrastive divergence. Specifically, we utilise the officially open-sourced implementation[6] of DULA to conduct contrastive divergence training. As depicted in Table 8, we find that energy discrepancy significantly outperforms contrastive divergence when employing a single MCMC step, and achieves performance comparable to CD-10. We attribute this superiority to the fact that CD-1 involves a biased estimation of the log-likelihood gradient due to inherent issues with non-convergent MCMC processes. In contrast, energy discrepancy does not suffer from this issue due to its consistent approximation.

The Efficacy of the Number of Negative Samples. In all experiments, we choose the number of negative samples as $M = 32$ irrespective of the dimension of the problem, to maximise computational efficiency within the constraints of our GPU capacity.

[6] https://github.com/ruqizhang/discrete-langevin

To investigate the impact of the number of negative samples on performance, we conduct experiments by training energy-based models on the static MNIST dataset with ED-Grid for different values of M. As detailed in Table 9, our results maintain comparable quality even as the number of negative samples is decreased. Notably, our approach offers greater parallelisation potential compared to the sequentially computed MCMC of contrastive divergence.

Table 9: Discrete image modelling results of ED-Grid on the static MNIST dataset with different M and $w = 1$.

	$M = 4$	$M = 8$	$M = 16$	$M = 32$
NLL	90.13	90.37	89.14	90.61

D.4 Training Ising Models

Task Descriptions. We further evaluate our methods for training the lattice Ising model, which has the form of

$$p(\mathbf{x}) \propto \exp(\mathbf{x}^T J \mathbf{x}),\ \mathbf{x} \in \{-1, 1\}^D,$$

where $J = \sigma A_D$ with $\sigma \in \mathbb{R}$ and A_D being the adjacency matrix of a $D \times D$ grid. Following Grathwohl et al. (2021); Zhang et al. (2022b,a), we generate training data through Gibbs sampling and use the generated data to fit a symmetric matrix J via energy discrepancy. Note that the training algorithms do not have access to the data-generating matrix J, only to the collection of samples.

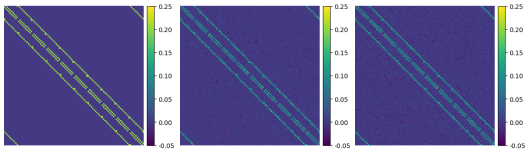


Figure 11: Results on learning Ising models. Left to right: ground truth, ED-Bern, ED-Grid.

Experimental Details. As in Grathwohl et al. (2021); Zhang et al. (2022a,b), we train a learnable connectivity matrix J_ϕ to estimate the true matrix J in the Ising model. To generate the training data, we simulate Gibbs sampling with $1,000,000$ steps for each instance to construct a dataset of $2,000$ samples. For energy discrepancy, we choose $w = 1, M = 32$ for all variants, $\epsilon = 0.1$ in Bernoulli perturbations. The parameter J_ϕ is learned by the Adam (Kingma & Ba, 2015) optimiser with a learning rate of 0.0001 and a batch size of 256. Following Zhang et al. (2022a), all models are trained with an l_1 regularisation with a coefficient in $\{100, 50, 10, 5, 1, 0.1, 0.01\}$ to encourage sparsity. The other setting is basically the same as Section F.2 in Grathwohl et al. (2021). We report the best result for each setting using the same hyperparameter searching protocol for all methods.

Qualitative Results. In Figure 11, we consider $D = 10 \times 10$ grids with $\sigma = 0.2$ and illustrate the learned matrix J using a heatmap. It can be seen that the variants of energy discrepancy can identify the pattern of the ground truth, confirming the effectiveness of our methods.

Quantitative Results. In the quantitative comparison to the baselines, we consider $D = 10 \times 10$ grids with $\sigma = 0.1, 0.2, \ldots, 0.5$ and $D = 9 \times 9$ grids with $\sigma = -0.1, -0.2$. The methods are evaluated by computing the negative log-RMSE between the estimated J_ϕ and the true matrix J. As shown in Table 10, our methods demonstrate comparable results to the baselines and, in certain settings, even outperform Gibbs and GWG, indicating that energy discrepancy is able to discover the underlying structure within the data.

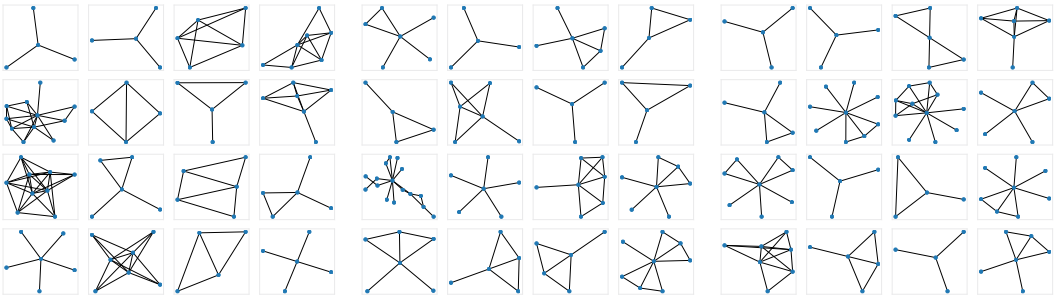
Table 10: Mean negative log-RMSE (higher is better) between the learned connectivity matrix J_ϕ and the true matrix J for different values of D and σ. The results of baselines are directly taken from Zhang et al. (2022a).

Method \ σ	$D = 10^2$					$D = 9^2$	
	0.1	0.2	0.3	0.4	0.5	−0.1	−0.2
Gibbs	4.8	4.7	**3.4**	**2.6**	**2.3**	4.8	4.7
GWG	4.8	4.7	**3.4**	**2.6**	**2.3**	4.8	4.7
EB-GFN	**6.1**	**5.1**	3.3	**2.6**	**2.3**	**5.7**	**5.1**
ED-Bern	5.1	4.0	2.9	2.5	**2.3**	5.1	4.3
ED-Grid	4.6	4.0	3.1	**2.6**	**2.3**	4.5	4.0

D.5 Graph Generation

Task Descriptions. The efficacy of our methods can be further demonstrated by producing high-quality graph samples. Following the setting in You et al. (2018), our model is evaluated on the Ego-small dataset, which comprises one-hop ego graphs extracted from the Citeseer network (Sen

https://doi.org/10.52202/079017-2516



(a) Training Data (b) ED-Bern (c) ED-Grid

Figure 12: Visualisation of the training data and samples drawn from the energy-based models learned by the variants of our approaches on the Ego-small dataset.

et al., 2008). We consider the following baselines[7] in graph generation, including GraphVAE (Simonovsky & Komodakis, 2018), DeepGMG (Li et al., 2018), GraphRNN (You et al., 2018), GNF (Liu et al., 2019), GrappAF (Shi et al., 2020), GraphDF (Luo et al., 2021), EDP-GNN (Niu et al., 2020), RMwGGIS (Liu et al., 2023), and contrastive divergence with GWG sampler (Grathwohl et al., 2021).

Experimental Details. Following the setup in You et al. (2018), we split the Ego-small dataset, allocating 80% for training and the remaining 20% for testing. To provide better insight into this task, we illustrate a subset of training data in Figure 12a. Notably, these training data examples closely resemble realistic one-hop ego graphs.

Table 11: Graph generation results in terms of MMD. *Avg.* denotes the average over three MMD results.

Method	*Degree*	*Cluster*	*Orbit*	*Avg.*
GraphVAE	0.130	0.170	0.050	0.117
DeepGMG	0.040	0.100	0.020	0.053
GraphRNN	0.090	0.220	0.003	0.104
GNF	0.030	0.100	0.001	0.044
GraphAF	0.030	0.110	0.001	0.047
GraphDF	0.040	0.130	0.010	0.060
EDP-GNN	0.052	0.093	0.007	0.050
EBM (GWG)	0.095	0.061	0.032	0.063
RMwGGIS	0.066	0.042	0.036	0.048
ED-Bern	0.063	0.054	0.014	0.044
ED-Grid	0.036	0.050	0.019	**0.035**

For a fair comparison, we parametrise the energy function via a 5-layer GCN (Kipf & Welling, 2017) with the ReLU activation and 16 hidden states for all energy-based approaches. For hyperparameters, we choose $M = 32, w = 1$ for all variants of energy discrepancy and $\epsilon = 0.1$ for the Bernoulli perturbation. Following the configuration in Liu et al. (2023), we apply the advanced version of RMwGGIS with the number of samples $s = 50$ (Liu et al., 2023, Equation 11). Regarding the EBM (GWG) baseline, we train it using persistent contrastive divergence with a buffer size of 200 samples and the MCMC steps being 50. To train the models, we use the Adam optimiser with a learning rate of 0.0001 and a batch size of 200. After training, we generate new graphs by first sampling N, which is the number of nodes to be generated, from the empirical distribution of the number of nodes in the training dataset, and then applying the GWG sampler (Grathwohl et al., 2021) with 50 MCMC steps from a randomly initialised Bernoulli noise. To assess the quality of these samples, we employ the MMD metric, evaluating it across three graph statistics, i.e., degrees, clustering coefficients, and orbit counts. Following the evaluation scheme in Liu et al. (2019), We trained 5 separate models of each type and performed 3 trials per model, then averaged the result over 15 runs.

Qualitative Results. We provide a visualisation of generated graphs from variants of our methods in Figures 12b and 12c. Notably, the majority of these generated graphs resemble one-hop ego graphs, illustrating their adherence to the graph characteristics in the training data.

Quantitative Results. In Table 11, we compare our methods to various baselines. It can be seen that our methods outperform most baselines in terms of the average of the three MMD metrics, indicating the faithful energy landscapes learned by the energy discrepancy approaches.

[7]There is insufficient information to reproduce EBM (GwG) and RMwGGIS precisely from Liu et al. (2023). We reran these two baselines with controlled hyperparameters for a fair comparison, while other baseline results were taken from their original papers.

E Naming Conventions and Parameters of Introduced Methods

This table summarises the naming conventions and available tuning parameters for all introduced methods. The structured perturbation TabED-Str uses different perturbations depending on the state space structure: On unstructured data, the uniform perturbation with tuning hyper-parameter t_{cat} is used, while on ordinally and cyclically structured data the ordinal perturbations and cyclical perturbations are used, respectively, with tuning parameter t_{base}.

Table 12: Overview of all introduced energy discrepancy methods

Name	Space (Discrete component)	Perturbation (Discrete component)	Tuning Parameter
ED-Bern	$\{0,1\}^d$	$\prod_{k=1}^d \mathrm{Bern}(\varepsilon)$	$\varepsilon = 0.5(1-e^{-2t})$
ED-Grid	$\{0,1\}^d$	$\sum_{k=1}^d \frac{1}{d}\delta_{\lvert y_k - x_k\rvert=1}$	None
TabED-Uni	$\otimes_{k=1}^d \{1,\dots,S_k\}$	$\prod_{k=1}^d \exp(tR^{\mathrm{unif}})_{y_k x_k}$ (Equation (7))	$t>0$
TabED-Grid	$\otimes_{k=1}^d \{1,\dots,S_k\}$	$\sum_{k=1}^d \frac{1}{d}\delta(y_k,\square)\delta(\mathbf{y}_{\neg k},\mathbf{x}_{\neg k})$	None
TabED-Cyc	$\otimes_{k=1}^d \{1,\dots,S_k\}$	$\prod_{k=1}^d \exp(t_k R^{\mathrm{cyc}})_{y_k x_k}$ (Proposition 1)	$t_k = S_k^2 t_{\mathrm{base}}$
TabED-Ord	$\otimes_{k=1}^d \{1,\dots,S_k\}$	$\prod_{k=1}^d \exp(t_k R^{\mathrm{ord}})_{y_k x_k}$ (Proposition 1)	$t_k = S_k^2 t_{\mathrm{base}}$
TabED-Str	$\otimes_{k=1}^d \{1,\dots,S_k\}$	$\prod_{k=1}^d \exp(t_k R^{k})_{y_k x_k}$ (Mixed)	Mixed

https://doi.org/10.52202/079017-2516

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our methodology describing the extension of energy discrepancy to discrete and mixed state spaces is given in Section 3 and Section 4, the proofs supporting our claims are found in section Appendix A of the appendix. We support our claims with experiments on discrete and mixed data in Section 6, demonstrating that the methodology generalises to settings of various types like categorical data, tabular data, and binary image data.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The main limitations are outlined in the conclusions section. To the best of our knowledge, the dominating factor in the performance of our method is the assumption that data can be assumed as independent samples of a positive density $p_{\mathrm{data}} > 0$. This assumption is violated for practically all data sets, but the extent to which the support of p_{data} is smaller than the state space deteriorates performance, either because of pre-mature convergence or because the perturbation does not explore the state space efficiently.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Proofs for our results can be found in Appendix A. Due to the nature of our results a proof sketch within the page limit was not feasible.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental details, regarding the network architecture and hyperparameters, are given in Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

https://doi.org/10.52202/079017-2516

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code is anonymised and given in the supplementary materials.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The full details are provided in Appendix D

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The quality of generated synthetic tabular data was tested by comparing the quality of a classifier trained on real and synthetic data. The error bar reflects the standard deviation of the classifier metric based on ten independently generated synthetic data sets sampled from the learned model. Error bars for negative log-likelihoods were not feasible with our computational resources due to the computational cost of annealed importance sampling.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The compute resources, as well as the required experimental runs, are detailed in Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors have reviewed the NeurIPS Code of Ethics and have considered the societal impact. The paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

https://doi.org/10.52202/079017-2516

Answer: [Yes]

Justification: The broader impact is discussed as part of the conclusion. Our method can be used for data imputation in tabular data which can be used downstream applications to discriminate or exclude if used irresponsibly, e.g. on biased data. However, we believe that our method is less prone to such applications than existing methods for data mining.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our paper demonstrates foundational research tested on publicly available datasets that were designed to test machine learning algorithms.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We properly acknowledge and cite all assets and resources used in the paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper introduces a training method for energy-based models and does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper introduces a training method for energy-based models and does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

https://doi.org/10.52202/079017-2516

Answer: [NA]

Justification: The paper introduces a training method for energy-based models and does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.