Training Discrete Energy-Based Models with Energy Discrepancy

Tobias Schröder[1]*, **Zijing Ou**[1]†, **Yingzhen Li**[1], **Andrew Duncan**[1,2]
[1] Imperial College London, UK, [2] The Alan Turing Institute, UK
{t.schroeder21, z.ou22, yingzhen.li, a.duncan}@imperial.ac.uk

Abstract

Training energy-based models (EBMs) on discrete spaces is challenging because sampling over such spaces can be difficult. We propose to train discrete EBMs with energy discrepancy (ED), a novel type of contrastive loss functional which only requires the evaluation of the energy function at data points and their perturbed counter parts, thus not relying on sampling strategies like Markov chain Monte Carlo (MCMC). Energy discrepancy offers theoretical guarantees for a broad class of perturbation processes of which we investigate three types: perturbations based on Bernoulli noise, based on deterministic transforms, and based on neighbourhood structures. We demonstrate their relative performance on lattice Ising models, binary synthetic data, and discrete image data sets.

1 Introduction

Building large-scale probabilistic models for discrete data is a critical challenge in machine learning for its broad applicability to perform inference and generation tasks on images, text, or graphs. Energy-based models (EBMs) are a class of particularly flexible models $p_{\mathrm{ebm}} \propto \exp(-U)$, where the modelling of the energy function U through a neural network function can be taylored to the data set of interest. However, EBMs are notoriously difficult to train due to the intractability of their normalisation.

The most popular paradigm for the training of EBMs is the contrastive divergence (CD) algorithm (Hinton, 2002) which performs approximate maximum likelihood estimation by using short-run Markov Chain Monte Carlo (MCMC) to approximate intractable expectations with respect to p_{ebm}. The success of CD has lead to rich research results on sampling from discrete distributions to enable fast and accurate estimation of the EBM (Zanella, 2020; Grathwohl et al., 2021; Zhang et al., 2022b; Sun et al., 2022b,a, 2023; Emami et al., 2023). However, training EBMs with CD remains challenging: Firstly, discrete probabilistic models often exhibit a large number of spurious modes which are difficult to explore even for the most advanced sampling algorithms. Secondly, CD lacks theoretical guarantees due to short run MCMC (Carreira-Perpinan & Hinton, 2005) and often times leads to malformed energy landscapes (Nijkamp et al., 2019).

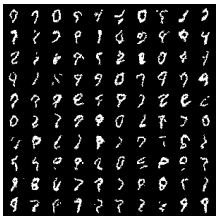

Figure 1: Generated samples from the EBM trained with Energy Discrepancy on static MNIST.

We propose the usage of a new type of loss function called Energy Discrepancy (ED) (Schröder et al., 2023) for the training of energy-based models on discrete spaces. The definition of ED only requires the evaluation of the EBM on positive and contrasting, negative samples. Unlike CD,

*Correspondence to: Tobias Schröder, t.schroeder21@imperial.ac.uk
†Code: https://github.com/J-zin/discrete-energy-discrepancy, z.ou22@imperial.ac.uk

ICML 2023 Workshop: Sampling and Optimization in Discrete Space (SODS 2023).

energy discrepancy does not require sampling from the model during training, thus allowing for fast training with theoretical guarantees. We demonstrate the effectiveness of ED by training Ising models, estimating discrete densities, and modelling discrete images in high-dimensions (see Figure 1 for an illustration).

2 Energy Discrepancies

Energy discrepancies are based on the idea that if information is processed through a channel $\mathcal{Q}$ then information will be lost. Mathematically, this is expressed through the data processing inequality $\mathrm{KL}(Qp_{\mathrm{data}} \parallel Qp_{\mathrm{ebm}}) \leq \mathrm{KL}(p_{\mathrm{data}} \parallel p_{\mathrm{ebm}})$. Consequently, the difference of the two KL divergences forms a valid loss for density estimation (Lyu, 2011). Retaining only terms that depend on the energy function U results in the energy discrepancy (Schröder et al., 2023):

Definition 1 (Energy Discrepancy)**.** *Let p_{data} be a positive density on a measure space $(\mathcal{X}, \mathrm{d}\mathbf{x})$*[3] *and let $q(\mathbf{y}|\mathbf{x})$ be a conditional probability density. Define the* contrastive potential *induced by q as*

$$U_q(\mathbf{y}) := -\log \sum_{\mathbf{x}' \in \mathcal{X}} q(\mathbf{y}|\mathbf{x}') \exp(-U(\mathbf{x}')) \tag{1}$$

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

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

The validity of this loss functional is given by the following non-parametric estimation result, previously stated in Schröder et al. (2023):

Theorem 1. *Let p_{data} be a positive probability density on $(\mathcal{X}, \mathrm{d}\mathbf{x})$. Assume that for all $\mathbf{x} \sim p_{\mathrm{data}}$ and $\mathbf{y} \sim q(\mathbf{y}|\mathbf{x})$, $\mathrm{Var}(\mathbf{x}|\mathbf{y}) > 0$. Then, the energy discrepancy ED_q is functionally convex in U and has, up to additive constants, a unique global minimiser $U^* = \operatorname{argmin} \mathrm{ED}_q(p_{\mathrm{data}}, U)$. Furthermore, this minimiser is the Gibbs potential for the data distribution, i.e. $p_{\mathrm{data}} \propto \exp(-U^*)$.*

We give the proof of Theorem 1 in Appendix A.1. The perturbation q can be chosen quite generally as long as it can be guaranteed that computing $\mathbf{y}$ comes at a loss of information which mathematically is expressed through the variance of recovering $\mathbf{x}$ from $\mathbf{y} \sim q(\mathbf{y}|\mathbf{x})$ being positive. In the next section, we propose some practical choices for q.

2.1 Training Discrete Energy-Based Models with Energy Discrepancy

The perturbation process q needs to be chosen under the following considerations: 1) The contrastive potential $U_q(\mathbf{y})$ has a numerically tractable approximation. 2) The negative samples obtained through q are informative for training the EBM when only finite amounts of data are available. We propose three categories for constructing perturbative processes:

Bernoulli Perturbation. For $\varepsilon \in (0, 1)$, let $\boldsymbol{\xi} \sim \mathrm{Bernoulli}(\varepsilon)^d$. On $\mathcal{X} = \{0, 1\}^d$, consider the perturbation $\mathbf{y} = \mathbf{x} + \boldsymbol{\xi} \bmod (2)$ which induces a symmetric transition density $q(\mathbf{y} - \mathbf{x})$ on $\{0, 1\}^d$. Due to the symmetry of q, we can then write the contrastive potential as

$$U_{\mathrm{bernoulli}}(\mathbf{y}) = -\log \sum_{\mathbf{x}' \in \mathcal{X}} q(\mathbf{y} - \mathbf{x}') \exp(-U(\mathbf{x}')) = -\log \mathbb{E}_{\mathbf{x}' \sim q(\mathbf{y} - \mathbf{x}')}[\exp(-U(\mathbf{x}'))]$$

The expectation on the right hand side can now be approximated via sampling M Bernoulli random variables $\boldsymbol{\xi}^j$ and taking the remainder of $(\mathbf{y} + \boldsymbol{\xi}^j)/2$. We denote this method as ED-Bern.

Deterministic Transformation. The perturbation q can also be defined through a deterministic information loosing map $g : \mathcal{X} \to \mathcal{Y}$, where the space $\mathcal{Y}$ may or may not be equal to $\mathcal{X}$ depending on the choice of g. The contrastive potential can be expressed in terms of the preimage of g, i.e.

$$U_g(\mathbf{y}) = -\log \sum_{\{\mathbf{x}' : g(\mathbf{x}') = \mathbf{y}\}} \exp(-U(\mathbf{x}')) = -\log \mathbb{E}_{\mathbf{x}' \sim \mathcal{U}(\{g^{-1}(\mathbf{y})\})}[\exp(-U(\mathbf{x}'))] - c$$

[3] 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

Table 1: Experiment results with seven 2D synthetic problems. We display the negative log-likelihood (NLL) and 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
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 (ours)	**20.039**	19.992	20.601	**19.710**	19.568	**20.084**	20.679
	ED-Pool (ours)	20.051	19.999	20.604	19.721	**19.531**	**20.084**	**20.676**
	ED-Grid (ours)	20.049	**19.965**	20.601	19.715	19.564	20.088	20.678
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 (ours)	0.120	0.014	0.137	**-0.088**	**0.046**	**0.045**	1.541
	ED-Pool (ours)	0.129	-0.003	**-0.021**	0.042	0.126	0.101	2.080
	ED-Grid (ours)	**0.097**	**-0.066**	0.022	0.018	0.351	0.097	2.049

with $c = \log |\{g^{-1}(\mathbf{y})\}|$. Again, the contrastive potential can be approximated through sampling M instances from the uniform distribution over the set $\{\mathbf{x}' : g(\mathbf{x}') = \mathbf{y}\}$. In our numerical experiments, we focus on the mean-pooling transform g_{pool} whose inverse are block-wise permutations. For details, see Appendix C.2. We denote this method as ED-Pool.

Neighbourhood-based Transformation. Finally, inspired from concrete score matching (Meng et al., 2022), we may define energy discrepancies based on neighbourhood maps $\mathbf{x} \mapsto \mathcal{N}(\mathbf{x}) \in \mathcal{X}^K$ which assign each point $\mathbf{x} \in \mathcal{X}$ a set of K neighbours[4]. We define the forward perturbation $q(\mathbf{y}|\mathbf{x})$ by selecting neighbours $\mathbf{y} \sim \mathcal{U}(\mathcal{N}(\mathbf{x}))$ uniformly at random. Conversely, the contrastive potential can be expressed in terms of the inverse neighbourhood $\mathbf{y} \mapsto \mathcal{N}^{-1}(\mathbf{y}) \in \mathcal{X}^K$, i.e. the set of points that have $\mathbf{y}$ to their neighbour. We then obtain for the contrastive potential

$$U_{\mathcal{N}}(\mathbf{y}) = -\log \frac{1}{K} \sum_{\mathbf{x}' \in \mathcal{X} : \mathbf{y} \in \mathcal{N}(\mathbf{x}')} \exp(-U(\mathbf{x}')) = -\log \mathbb{E}_{\mathbf{x}' \sim \mathcal{U}(\{\mathcal{N}^{-1}(\mathbf{y})\})}[\exp(-U(\mathbf{x}'))] \,.$$

In practice, we choose the grid neighbourhood (Appendix C.3) and denote this method by ED-Grid.

Stabilising Training. Above schemes permit the approximation of the contrastive potential from M samples which are generated by first sampling $\mathbf{y} \sim q(\mathbf{y}|\mathbf{x})$, after which we compute M approximate recoveries $\mathbf{x}_-^j$. The full loss can then be constructed for each data point $\mathbf{x}_+ \sim p_{\text{data}}$ by calculating $\log \sum_{j=1}^M \exp(U(\mathbf{x}_+) - U(\mathbf{x}_-^j)) - \log(M)$ using the numerically stabilised logsumexp function. In practice, however, we find that this estimator for energy discrepancy is biased due to the logarithm and can exhibit high variance. To stabilise training, we introduce an offset for the logarithm which introduces a deterministic lower bound for the loss. This yields the energy discrepancy loss function

$$\mathcal{L}_{q,M,w}(U) := \frac{1}{N} \sum_{i=1}^N \log \left(w + \sum_{j=1}^M \exp(U(\mathbf{x}_+^i) - U(\mathbf{x}_-^{i,j})) \right) - \log(M) \tag{2}$$

with $\mathbf{x}_+^i \sim p_{\text{data}}$. In Appendix C.5 we proof that this approximation is consistent for any fixed w:

Theorem 2. *For every $\varepsilon > 0$ there exist $N, M \in \mathbb{N}$ such that $|\mathcal{L}_{q,M,w}(U) - \text{ED}_q(p_{\text{data}}, U)| < \varepsilon$ a.s..*

3 Experiments

Training Ising Models. We evaluate the proposed methods on 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 $D \times D$

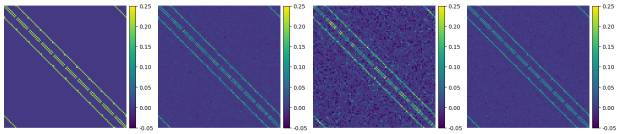

Figure 2: Experiment results on learning lattice Ising models. Left to right: ground truth, ED-Bern, ED-Pool, ED-Grid.

[4]We are making the assumption that the numbers of neighbours is the same for each point. A more general case is discussed in Appendix C.4.

Table 2: Experimental results on 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 (ours)	ED-Pool (ours)	ED-Grid (ours)
Static MNIST	117.17	**80.01**	102.43	80.71	95.38	168.07	90.15
Dynamic MNIST	121.19	**80.51**	105.75	81.29	97.03	144.26	81.26
Omniglot	142.06	94.72	112.59	145.68	97.87	118.66	**94.64**
Caltech Silhouettes	163.50	**96.20**	185.57	100.52	96.36	501.96	117.70

grid. Following Zhang et al. (2022a), we generate training data through Gibbs sampling and use the generated data to fit a symmetric matrix J via energy discrepancy. In Figure 2, 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. We defer experimental details and quantitative results comparing with baselines to Appendix E.1.

Discrete Density Estimation. In this experiment, we follow the experimental setting of Dai et al. (2020); Zhang et al. (2022a), which aims to model discrete densities over 32-dimensional binary data that are discretisations of continuous densities on the plane (see Figure 4). Specifically, we convert each planar data point $\hat{\mathbf{x}} \in \mathbb{R}^2$ to a binary data point $\mathbf{x} \in \{0, 1\}^{32}$ via Gray code (Gray, 1953). Consequently, the models face the challenge of modeling data in a discrete space, which is particularly difficult due to the non-linear transformation from $\hat{\mathbf{x}}$ to $\mathbf{x}$.

We compare our methods to three baselines: PCD (Tieleman, 2008), ALOE+ (Dai et al., 2020), and EB-GFN (Zhang et al., 2022a). The experimental details are given in Appendix E.2. For qualitative evaluation, we visualise the energy landscapes learned by our methods in Figure 3. It shows that energy discrepancy is able to faithfully model multi-modal distributions and accurately learn the sharp edges present in the data support. For further qualitative comparisons, we refer to the energy landscapes of baseline methods presented in Figure C.2 of Zhang et al. (2022a). Moreover, we quantitatively evaluate different methods in Table 1 by showing the negative log-likelihood (NLL) and the exponential Hamming MMD (Gretton et al., 2012). Perhaps surprisingly, we find that energy discrepancy outperforms the baselines on most settings, despite not requiring MCMC simulation like PCD or training an additional variational network like ALOE and EB-GFN. A possible explanation for this are biases introduced by short-run MCMC sampling in the case of PCD or non-converged variational proposals in ALOE. By definition, ED transforms the data distribution as well as the energy function which corrects for such biases.

Discrete Image Modelling. Here, we evaluate our methods in discrete high-dimensional spaces. Following the settings in Grathwohl et al. (2021); Zhang et al. (2022b), we conduct experiments on four different binary image datasets. Training details are given in Appendix E.3. After training, we adopt Annealed Importance Sampling (Neal, 2001) to estimate the log-likelihoood.

The baselines include persistent contrastive divergence with 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 NLLs on the test set are reported in Table 2. We see that energy discrepancy yields comparable performances to the baselines, while ED-Pool is unable to capture the data distribution. We emphasise that energy discrepancy only requires M (here, $M = 32$) evaluations of the energy function per data point in parallel. This is notably fewer than contrastive divergence, which requires simulating multiple MCMC steps without parallelisation. We also visualise the generated samples in Figure 11, which showcase the diversity and high quality of the images generated by ED-Bern and ED-Grid. However, we observed that ED-Pool suffers from mode collapse.

4 Conclusion and Outlook

In this paper we demonstrate how energy discrepancy can be used for efficient and competitive training of energy-based models on discrete data without MCMC. The loss can be defined based on a large class of perturbative processes of which we introduce three types: noise, determinstic transform, and neighbourhood-based transform. Our results show that the choice of perturbation matters and motivates further research on effective choices depending on the data structure of interest.

We observe empirically that similarly to other contrastive losses, energy discrepancy shows limitations when the ambient dimension of $\mathcal{X}$ is significantly larger than the intrinsic dimension of the data. In these cases, training is aided significantly by a base distribution that models the lower-dimensional space populated by data. For this reason, the adoption of ED on new data sets or different data structures may require adjustments to the methodology such as learning appropriate base distributions and finding more informative perturbative transforms.

For future work, we are interested in how this work extends to highly structured data such as graphs or text. These settings may require a deeper understanding of how the perturbation influences the performance of ED and what is gained from gradient information in CD (Zhang et al., 2022b; Grathwohl et al., 2021) or ratio matching (Liu et al., 2023).

Acknowledgements

TS would like to thank G.A. Pavliotis for insightful discussions leading up to the presented work. 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. ABD was supported by Wave 1 of The UKRI Strategic Priorities Fund under the EPSRC Grant EP/T001569/1 and EPSRC Grant EP/W006022/1, particularly the "Ecosystems of Digital Twins" theme within those grants and The Alan Turing Institute. We thank the anonymous reviewer for their comments.

References

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 Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 726–734. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/ceylan18a.html.

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.

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): 025014, 2023.

Grathwohl, W., Swersky, K., Hashemi, M., Duvenaud, D., and Maddison, C. J. Oops I took a gradient: Scalable sampling for discrete distributions. *arXiv preprint arXiv:2102.04509*, 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):723–773, 2012.

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*, pp. 770–778, 2016.

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

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

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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.

Liu, M., Liu, H., and Ji, S. Gradient-guided importance sampling for learning binary energy-based models. 2023.

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. URL https://proceedings.neurips.cc/paper/2011/file/a3f390d88e4c41f2747bfa2f1b5f87db-Paper.pdf.

Lyu, S. Interpretation and generalization of score matching. *arXiv preprint arXiv:1205.2629*, 2012.

Majerek, D., Nowak, W., and Ziba, W. Conditional strong law of large number. *International Journal of Pure and Applied Mathematics*, 20, 01 2005.

Meng, C., Choi, K., Song, J., and Ermon, S. Concrete score matching: Generalized score matching for discrete data. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 34532–34545. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/df04a35d907e894d59d4eab1f92bc87b-Paper-Conference.pdf.

Neal, R. M. Annealed importance sampling. *Statistics and computing*, 11:125–139, 2001.

Nijkamp, E., Hill, M., Zhu, S.-C., and Wu, Y. N. Learning non-convergent non-persistent short-run MCMC toward energy-based model. *arXiv preprint arXiv:1904.09770*, 2019.

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.

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

Rhodes, B. and Gutmann, M. Enhanced gradient-based MCMC in discrete spaces. *arXiv preprint arXiv:2208.00040*, 2022.

Schröder, T., Ou, Z., Lim, J. N., Li, Y., Vollmer, S. J., and Duncan, A. B. Energy discrepancies: A score-independent loss for energy-based models, 2023. URL https://arxiv.org/abs/2307.06431.

Sun, H., Dai, H., and Schuurmans, D. Optimal scaling for locally balanced proposals in discrete spaces. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23867–23880. Curran Associates, Inc., 2022a. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/96c6f409a374b5c81d2efa4bc5526f27-Paper-Conference.pdf.

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*, pp. 6290–6313. PMLR, 2023.

Tieleman, T. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th international conference on Machine learning*, pp. 1064–1071, 2008.

van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL http://arxiv.org/abs/1807.03748.

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. *arXiv preprint arXiv:2202.01361*, 2022a.

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 "Training Discrete EBMs with Energy Discrepancy"

Contents

A Abstract Proofs and Derivations

A.1 Proof of the Non-Parametric Estimation Theorem 1

In this subsection we give a formal proof for the uniqueness of minima of $\mathrm{ED}_q(p_{\mathrm{data}}, U)$ as a functional in the energy function U. We first reiterate the theorem as stated in the paper:

Theorem 1. *Let p_{data} be a positive probability density on $(\mathcal{X}, \mathrm{d}\mathbf{x})$. Assume that for all $\mathbf{x} \sim p_{\mathrm{data}}$ and $\mathbf{y} \sim q(\mathbf{y}|\mathbf{x})$, $\mathrm{Var}(\mathbf{x}|\mathbf{y}) > 0$. Then, the energy discrepancy ED_q is functionally convex in U and has, up to additive constants, a unique global minimiser $U^* = \operatorname{argmin} \mathrm{ED}_q(p_{\mathrm{data}}, U)$. Furthermore, this minimiser is the Gibbs potential for the data distribution, i.e. $p_{\mathrm{data}} \propto \exp(-U^*)$.*

We test energy discrepancy on the first and second order optimality conditions, i.e. we test that the first functional derivative of ED vanishes in U^* and that the second functional derivative is positive definite. For uniqueness and well-definedness, we constrain the optimisation domain to the following set:

$$\mathcal{G} := \left\{ U : \mathcal{X} \mapsto \mathbb{R} \text{ such that } \exp(-U) \in L^1(\mathcal{X}, \mathrm{d}\mathbf{x}), \; U \in L^1(p_{\mathrm{data}}), \text{ and } \min_{\mathbf{x} \in \mathcal{X}} U(x) = 0 \right\}$$

and require that there exists a $U^* \in \mathcal{G}$ such that $\exp(-U^*) \propto p_{\mathrm{data}}$. We now start with the following lemmata and then complete the proof of Theorem 1 in Corollary 1.

Lemma 1. *Let $h \in \mathcal{G}$ be arbitrary. The first variation of ED_q is given by*

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon} \mathrm{ED}_q(p_{\mathrm{data}}, U + \epsilon h)\Big|_{\epsilon=0} = \mathbb{E}_{p_{\mathrm{data}}(\mathbf{x})}[h(\mathbf{x})] - \mathbb{E}_{p_{\mathrm{data}}(\mathbf{x})}\mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\mathbb{E}_{p_U(\mathbf{z}|\mathbf{y})}[h(\mathbf{z})] \tag{3}$$

where $p_U(\mathbf{z}|\mathbf{y}) = \frac{q(\mathbf{y}|\mathbf{z}) \exp(-U(\mathbf{z}))}{\sum_{\mathbf{z}' \in \mathcal{X}} q(\mathbf{y}|\mathbf{z}') \exp(-U(\mathbf{z}'))}$.

Proof. We define the short-hand notation $U_\epsilon := U + \epsilon h$. The energy discrepancy at U_ε reads

$$\mathrm{ED}_q(p_{\mathrm{data}}, U_\epsilon) = \mathbb{E}_{p_{\mathrm{data}}(\mathbf{x})}[U_\epsilon(\mathbf{x})] + \mathbb{E}_{p_{\mathrm{data}}(\mathbf{x})}\mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log \sum_{\mathbf{z}\in\mathcal{X}} q(\mathbf{y}|\mathbf{z}) \exp(-U_\epsilon(\mathbf{z}))\right].$$

For the first functional derivative, we only need to calculate

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon} \log \sum_{\mathbf{z}\in\mathcal{X}} q(\mathbf{y}|\mathbf{z}) \exp(-U_\epsilon(\mathbf{z})) = \sum_{\mathbf{z}\in\mathcal{X}} \frac{-q(\mathbf{y}|\mathbf{z})h(\mathbf{z})\exp(-U_\epsilon(\mathbf{z}))}{\sum_{\mathbf{z}'\in\mathcal{X}} q(\mathbf{y}|\mathbf{z}')\exp(-U_\epsilon(\mathbf{z}'))} = -\mathbb{E}_{p_{U_\epsilon}(\mathbf{z}|\mathbf{y})}[h(\mathbf{z})]. \quad (4)$$

Plugging this expression into $\mathrm{ED}_q(p_{\mathrm{data}}, U_\epsilon)$ and setting $\epsilon = 0$ yields the first variation of ED_q. □

Lemma 2. *The second variation of* ED_q *is given by*

$$\frac{\mathrm{d}^2}{\mathrm{d}\epsilon^2}\mathrm{ED}_q(p_{\mathrm{data}}, U + \epsilon h)\Big|_{\epsilon=0} = \mathbb{E}_{p_{\mathrm{data}}(\mathbf{x})}\mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\mathrm{Var}_{p_U(\mathbf{z}|\mathbf{y})}[h(\mathbf{z})].$$

Proof. For the second order term, we have based on equation 4 and the quotient rule for derivatives:

$$\begin{aligned}
&\frac{\mathrm{d}^2}{\mathrm{d}\epsilon^2} \log \sum_{\mathbf{z}\in\mathcal{X}} q(\mathbf{y}|\mathbf{z}) \exp(-U_\epsilon(\mathbf{z})) \\
&\quad= \frac{\sum_{\mathbf{z}\in\mathcal{X}} q(\mathbf{y}|\mathbf{z})\exp(U_\epsilon(\mathbf{z}))\, h^2(\mathbf{z}) \;\sum_{\mathbf{z}'\in\mathcal{X}} q(\mathbf{y}|\mathbf{z}')\exp(-U_\epsilon(\mathbf{z}'))}{\left(\sum_{\mathbf{z}'\in\mathcal{X}} q(\mathbf{y}|\mathbf{z}')\exp(-U_\epsilon(\mathbf{z}'))\right)^2} \\
&\qquad- \frac{\sum_{\mathbf{z}\in\mathcal{X}} q(\mathbf{y}|\mathbf{z})\exp(U_\epsilon(\mathbf{z}))h(\mathbf{z}) \sum_{\mathbf{z}'\in\mathcal{X}} q(\mathbf{y}|\mathbf{z}')\exp(-U_\epsilon(\mathbf{z}'))h(\mathbf{z}')}{\left(\sum_{\mathbf{z}'\in\mathcal{X}} q(\mathbf{y}|\mathbf{z}')\exp(-U_\epsilon(\mathbf{z}'))\right)^2} \\
&\quad= \mathbb{E}_{p_{U_\epsilon}(\mathbf{z}|\mathbf{y})}[h^2(\mathbf{z})] - \mathbb{E}_{p_{U_\epsilon}(\mathbf{z}|\mathbf{y})}[h(\mathbf{z})]^2 = \mathrm{Var}_{p_{U_\epsilon}(\mathbf{z}|\mathbf{y})}[h(\mathbf{z})]\,.
\end{aligned}$$

We obtain the desired result by interchanging the outer expectations with the derivatives in ϵ. □

Corollary 1. *Let* $c = \min_{\mathbf{x}\in\mathcal{X}}(-\log p_{\mathrm{data}}(\mathbf{x}))$. *For* $U^* = -\log(p_{\mathrm{data}}) - c \in \mathcal{G}$ *it holds that*

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\epsilon}\mathrm{ED}_q(p_{\mathrm{data}}, U^* + \epsilon h)\Big|_{\epsilon=0} = 0 \\
&\frac{\mathrm{d}^2}{\mathrm{d}\epsilon^2}\mathrm{ED}_q(p_{\mathrm{data}}, U^* + \epsilon h)\Big|_{\epsilon=0} > 0 \qquad \textit{for all} \qquad h\,,
\end{aligned}$$

Furthermore, U^* *is the unique global minimiser of* $\mathrm{ED}_q(p_{\mathrm{data}}, \cdot)$ *in* $\mathcal{G}$.

Proof. By definition, the variance is non-negative, i.e. for every $h \in \mathcal{G}$:

$$\frac{\mathrm{d}^2}{\mathrm{d}\epsilon^2}\mathrm{ED}_q(p_{\mathrm{data}}, U + \epsilon h)\Big|_{\epsilon=0} = \mathrm{Var}_{p_U(\mathbf{z}|\mathbf{y})}[h(\mathbf{z})] \geq 0\,.$$

Consequently, the energy discrepancy is convex and an extremal point of $\mathrm{ED}_q(p_{\mathrm{data}}, \cdot)$ is a global minimiser. We are left to show that the minimiser is obtained at U^* and unique. First of all, we have for U^*:

$$\begin{aligned}
\mathbb{E}_{p_{U^*}(\mathbf{z}|\mathbf{y})}[h(\mathbf{z})] &= \sum_{\mathbf{z}\in\mathcal{X}} \frac{q(\mathbf{y}|\mathbf{z})\exp(-U^*(\mathbf{z}))}{\sum_{\mathbf{z}'\in\mathcal{X}} q(\mathbf{y}|\mathbf{z}')\exp(-U^*(\mathbf{z}'))} h(\mathbf{z}) \\
&= \sum_{\mathbf{z}\in\mathcal{X}} \frac{q(\mathbf{y}|\mathbf{z})p_{\mathrm{data}}(\mathbf{z})}{\sum_{\mathbf{z}'\in\mathcal{X}} q(\mathbf{y}|\mathbf{z}')p_{\mathrm{data}}(\mathbf{z}')} h(\mathbf{z}).
\end{aligned}$$

By applying the outer expectations we obtain

$$\begin{aligned}
\mathbb{E}_{p_{\mathrm{data}}(\mathbf{x})}\mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\mathbb{E}_{p_{U^*}(\mathbf{z}|\mathbf{y})}[h(\mathbf{z})] &= \sum_{\mathbf{x}\in\mathcal{X}} p_{\mathrm{data}}(\mathbf{x}) \sum_{\mathbf{y}\in\mathcal{Y}} \left(q(\mathbf{y}|\mathbf{x}) \sum_{z\in\mathcal{X}} \left(\frac{q(\mathbf{y}|\mathbf{z})p_{\mathrm{data}}(\mathbf{z})}{\sum_{\mathbf{z}'\in\mathcal{X}} q(\mathbf{y}|\mathbf{z}')p_{\mathrm{data}}(\mathbf{z}')} h(\mathbf{z}) \right) \right) \\
&= \sum_{\mathbf{z}\in\mathcal{X}} \sum_{\mathbf{y}\in\mathcal{Y}} q(\mathbf{y}|\mathbf{x})p_{\mathrm{data}}(\mathbf{z})h(\mathbf{z}) \\
&= \mathbb{E}_{p_{\mathrm{data}}(\mathbf{z})}[h(\mathbf{z})],
\end{aligned}$$

where we used that the marginal distributions $\sum_{\mathbf{x}\in\mathcal{X}} p_{\text{data}}(\mathbf{x})q(\mathbf{y}|\mathbf{x})$ cancel out and the conditional probability density integrates to one. This implies

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon}\mathrm{ED}_q(p_{\text{data}}, U^* + \epsilon h)\Big|_{\epsilon=0} = \mathbb{E}_{p_{\text{data}}(\mathbf{z})}[h(\mathbf{z})] - \mathbb{E}_{p_{\text{data}}(\mathbf{z})}[h(\mathbf{z})] = 0.$$

for all $h \in \mathcal{G}$. We now show that

$$\frac{\mathrm{d}^2}{\mathrm{d}\epsilon^2}\mathrm{ED}_q(p_{\text{data}}, U^* + \epsilon h)\Big|_{\epsilon=0} = \mathbb{E}_{p_{\text{data}}(\mathbf{x})}\mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\mathrm{Var}_{p_{\text{data}}(\mathbf{z}|\mathbf{y})}[h(\mathbf{z})] > 0\,.$$

Assume that the second variation was zero. Since the perturbed data distribution $\sum_{\mathbf{x}\in\mathcal{X}} p_{\text{data}}(\mathbf{x})q(\mathbf{y}|\mathbf{x})$ is positive, the second variation at U^* is zero if and only if the conditional variance $\mathrm{Var}_{p_{\text{data}}(\mathbf{z}|\mathbf{y})}[h(\mathbf{z})] = 0$. Since $U^* + \varepsilon h \in \mathcal{G}$, the function h can not be constant. By definition of the conditional variance, $h(\mathbf{z})$ must then be a deterministic function of $\mathbf{y} \sim \sum_{\mathbf{x}\in\mathcal{X}} q(\mathbf{y}|\mathbf{x})p_{\text{data}}(\mathbf{x})$. Since h was arbitrary, there exists a measurable map g such that $\mathbf{z} = g(\mathbf{y})$ and $\mathrm{Var}_{p_{\text{data}}(\mathbf{z}|\mathbf{y})}[\mathbf{z}] = 0$ which is a contradiction to our assumptions. Consequently, U^* is the unique global minimiser of ED_q which completes the statement in Theorem 1. □

B Connections to other Methods

In this section, we follow Schröder et al. (2023).

B.1 Connections of Energy Discrepancy with Contrastive Divergence

The contrastive divergence update can be derived from an energy discrepancy when, for E_θ fixed, q satisfies the detailed balance relation

$$q(\mathbf{y}|\mathbf{x})\exp(-E_\theta(\mathbf{x})) = q(\mathbf{x}|\mathbf{y})\exp(-E_\theta(\mathbf{y}))\,.$$

To see this, we calculate the contrastive potential induced by q: We have

$$-\log\sum_{\mathbf{x}'\in\mathcal{X}} q(\mathbf{y}|\mathbf{x}')\exp(-E_\theta(\mathbf{x}')) = -\log\sum_{\mathbf{x}'\in\mathcal{X}} q(\mathbf{x}'|\mathbf{y})\exp(-E_\theta(\mathbf{y})) = E_\theta(\mathbf{y})\,.$$

Consequently, the energy discrepancy induced by q is given by

$$\mathrm{ED}_q(p_{\text{data}}, E_\theta) = \mathbb{E}_{p_{\text{data}}(\mathbf{x})}[E_\theta(\mathbf{x})] - \mathbb{E}_{p_{\text{data}}(\mathbf{x})}\mathbb{E}_{q(\mathbf{y}|\mathbf{x})}[E_\theta(\mathbf{y})]\,.$$

Updating θ based on a sample approximation of this loss leads to the contrastive divergence update

$$\Delta\theta \propto \frac{1}{N}\sum_{i=1}^N \nabla_\theta E_\theta(\mathbf{x}^i) - \frac{1}{N}\sum_{i=1}^N \nabla_\theta E_\theta(\mathbf{y}^i) \quad \mathbf{y}^i \sim q(\cdot|\mathbf{x}^i)$$

It is important to notice that the distribution q depends on E_θ and needs to adjusted in each step of the algorithm. For fixed q, $\mathrm{ED}_q(p_{\text{data}}, E_\theta)$ satisfies Theorem 1. This means that each step of contrastive divergence optimises a loss with minimiser $E_\theta^* = -\log p_{\text{data}} + c$. However, the loss function changes in each step of contrastive divergence. The connection also highlights the importance Metropolis-Hastings adjustment to ensure that the implied q distribution satisfies the detailed balance relation.

B.2 Derivation of Energy Discrepancy from KL Contractions

A Kullback-Leibler contraction is the divergence function $\mathrm{KL}(p_{\text{data}} \parallel p_{\text{ebm}}) - \mathrm{KL}(Qp_{\text{data}} \parallel Qp_{\text{ebm}})$ (Lyu, 2011) for the convolution operator $Qp(\mathbf{y}) = \sum_{\mathbf{x}'\in\mathcal{X}} q(\mathbf{y}|\mathbf{x}')p(\mathbf{x}')$. The linearity of the convolution operator retains the normalisation of the measure, i.e. for the energy-based distribution p_{ebm} we have

$$Qp_{\text{ebm}} = \frac{1}{Z_U}\sum_{\mathbf{x}'\in\mathcal{X}} q(\mathbf{y}|\mathbf{x}')\exp(-U(\mathbf{x}')) \quad \text{with} \quad Z_U = \sum_{\mathbf{x}'\in\mathcal{X}}\exp(-U(\mathbf{x}'))\,.$$

The KL divergences then become with $U_q := -\log Q\exp(-U(\mathbf{x}))$

$$\mathrm{KL}(p_{\text{data}} \parallel p_{\text{ebm}}) = \mathbb{E}_{p_{\text{data}}(\mathbf{x})}[\log p_{\text{data}}(\mathbf{x})] + \mathbb{E}_{p_{\text{data}}(\mathbf{x})}[U(\mathbf{x})] + \log Z_U$$

$$\mathrm{KL}(Qp_{\text{data}} \parallel Qp_{\text{ebm}}) = \mathbb{E}_{Qp_{\text{data}}(\mathbf{y})}[\log Qp_{\text{data}}(\mathbf{y})] + \mathbb{E}_{Qp_{\text{data}}(\mathbf{y})}[U_q(\mathbf{y})] + \log Z_U$$

Since the normalisation cancels when subtracting the two terms we find

$$\mathrm{KL}(p_{\text{data}} \parallel p_{\text{ebm}}) - \mathrm{KL}(Qp_{\text{data}} \parallel Qp_{\text{ebm}}) = \mathrm{ED}_q(p_{\text{data}}, U) + c$$

where c is a constant that contains the U-independent entropies of p_{data} and Qp_{data}.

C Sample Approximations of Energy Discrepancies

In this section, we discuss practical implementations of the mean-pooling transform as an information destroying deterministic process and the grid-neighbourhood as a neighbourhood-based transformation.

C.1 General Strategy

As a general strategy, the contrastive potential has to be written as an expectation over an appropriate to be determined distribution $p_{\mathrm{neg},q,\mathbf{y}}$ that depends on the chosen perturbation process and on the point where the contrastive potential is evaluated, i.e.

$$U_q(\mathbf{y}) = -\log \mathbb{E}_{p_{\mathrm{neg},q,\mathbf{y}}(\mathbf{x}')} \exp(-U(\mathbf{x}')) \tag{5}$$

which allows the evaluation of the contrastive potential via sampling from $p_{\mathrm{neg},q,\mathbf{y}}$. The energy discrepancy can then be written as

$$\mathrm{ED}_q(p_{\mathrm{data}}, U) = \mathbb{E}_{p_{\mathrm{data}}(\mathbf{x})}\mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log \mathbb{E}_{p_{\mathrm{neg},q,\mathbf{y}}(\mathbf{x}')}\left[\exp(U(\mathbf{x}) - U(\mathbf{x}'))\right]\right] \tag{6}$$

by using properties of the logarithm and exponential and the fact that $U(\mathbf{x})$ does not depend on the expectations taken in $\mathbf{y}$ and $\mathbf{x}'$. The loss can then be approximated via ancestral sampling. We first sample a batch $\mathbf{x}_+^i \sim p_{\mathrm{data}}$, subsequently sample its perturbed counter part $\mathbf{y}^i \sim q(\cdot|\mathbf{x}_+^i)$, and finally sample M negative samples $\mathbf{x}_-^{i,j} \sim p_{\mathrm{neg},q,\mathbf{y}^i}$. Sometimes, the perturbed sample $\mathbf{y}^i$ is never explicitely computed in the process. As described in Equation (2), the approximation is always stabilised through tunable hyper-parameter w which finally yields the loss function

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

The justification for the stabilisation is two-fold. Firstly, the logarithm makes the Monte-Carlo approximation of the contrastive potential biased due to Jensens inequality. The bias is negative, given to leading order by the variance of the approximation, and depends on the energy function U. Thus, the optimiser may start to optimise for a high bias and high variance estimator of the contrastive potential rather than learning the data distribution. While this issue can be alleviated by significantly large choices for M, it is much more practical to introduce a deterministic lower bound to the loss-functional through the stabilisation w, which prevents the bias and logarithm from diverging. Secondly, the effect of the stabilisation goes to zero as M increases. Thus, the asymptotic limit for M and N large is retained through the stabilisation. For more details and analogous arguments in the continuous case, see Schröder et al. (2023).

C.2 Mean Pooling Transform

We describe the mean-pooling transform on the example of image data which takes values in the space $\{0,1\}^{h\times w}$. We fix a window size s and reshape each data-point into blocks of size $s \times s$, i.e.

$$\{0,1\}^{h\times w} \to \{0,1\}^{s\times s\times \frac{h}{s}\times\frac{w}{s}}, \quad \mathbf{x} \mapsto \bar{\mathbf{x}}$$

The mean pooling transform g_{pool} computes the average over each block $\bar{\mathbf{x}}_{\bullet,\bullet,i,j}$ for $i = 1, 2, \ldots, h/s$ and $j = 1, 2, \ldots, w/s$. The corresponding preimage of the mean pooling transform is given by the set of points which are identical to $\mathbf{x}$ up to block-wise permutation, i.e.

$$g^{-1}(g_{\mathrm{pool}}(\mathbf{x})) = \{\mathbf{x}' \in \mathcal{X} : \text{ there exist } \pi_{i,j} \in S_{s\times s} \text{ s.t. } \bar{\mathbf{x}}'_{l,k,i,j} = \bar{\mathbf{x}}'_{\pi_{i,j}(l,k),i,j} \text{ for all } l,k,i,j\}$$

where $S_{s\times s}$ denotes the permutation group for matrices of size $s \times s$. In practice, the mean-pooled data point has to never be computed, only the block wise permutations of the data point are required. Consequently, we obtain negative samples through $\mathbf{x}_-^{i,j} \sim \mathcal{U}(g^{-1}(g_{\mathrm{pool}}(\mathbf{x}^i)))$, i.e. via block wise permutation of the entries of each data point $\mathbf{x}^i$.

Strictly speaking, this transformation violates the assumptions of Theorem 1 for data points that only consist of blocks that average to 1 or 0. Since this is only the case for a small set of the state space, we assume this violation to be negligible.

C.3 Grid Neighborhood

The grid neighbourhood for $\mathbf{x} \in \{0,1\}^d$ is constructed as

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

where $\mathbf{e}_k$ is a vector of zeros with a one in the k-th entry. This neighbourhood structure is symmetric, i.e. $\mathcal{N}_{\text{grid}}^{-1}(\mathbf{y}) = \mathcal{N}_{\text{grid}}(\mathbf{y})$. Consequently, the negative samples are created by sampling from

$$\mathbf{x}_-^{i,j} \sim \mathcal{U}(\mathcal{N}_{\text{grid}}(\mathbf{y}^i)) \quad \text{with} \quad \mathbf{y}^i \sim \mathcal{U}(\mathcal{N}_{\text{grid}}(\mathbf{x}^i))$$

Notice that each negative sample is the second neighbour of the positive sample, and with a small chance the positive sample itself.

C.4 Directed Neighbourhood Structures

More generally, the neighbourhood structure may form a non-symmetric directed graph for which the neighbourhood maps $\mathcal{N}^{-1}$ and $\mathcal{N}$ don't coincide. In this case, an additional weighting-term is introduced. We denote the number of neighbours of $\mathbf{x}$ as $K_{\mathbf{x}} = |\mathcal{N}(\mathbf{x})|$ and the number of elements of which $\mathbf{y}$ is a neighbour as $K'_{\mathbf{y}} = |\mathcal{N}^{-1}(\mathbf{y})|$. The forward transition density is given by the uniform distribution, i.e.

$$q(\mathbf{y}|\mathbf{x}) = \begin{cases} 1/K_{\mathbf{x}} & \text{if} \quad \mathbf{y} \in \mathcal{N}(\mathbf{x}) \\ 0 & \text{else} \end{cases} \tag{7}$$

We then have

$$\begin{aligned} U_{\mathcal{N}}(\mathbf{y}) &= \log \sum_{\mathbf{x}' \in \mathcal{X}} q(\mathbf{y}|\mathbf{x}') \exp(-U(\mathbf{x}')) \\ &= \log \sum_{\mathbf{x}' \in \mathcal{N}^{-1}(\mathbf{y})} \frac{1}{K_{\mathbf{x}'}} \exp(-U(\mathbf{x}')) \\ &= \log \frac{1}{K'_{\mathbf{y}}} \sum_{\mathbf{x}' \in \mathcal{N}^{-1}(\mathbf{y})} \frac{K'_{\mathbf{y}}}{K_{\mathbf{x}'}} \exp(-U(\mathbf{x}')) \\ &= \log \mathbb{E}_{\mathbf{x}' \sim \mathcal{U}(\{\mathcal{N}^{-1}(\mathbf{y})\})} [\omega_{\mathbf{y}\mathbf{x}'} \exp(-U(\mathbf{x}'))] \end{aligned}$$

where we introduced the weighting term $\omega_{\mathbf{y}\mathbf{x}'} = K'_{\mathbf{y}}/K_{\mathbf{x}'}$.

C.5 Consistency of our Approximation

The following proof is similar to Schröder et al. (2023). We first restate the consistency result:

Theorem 2. *For every $\varepsilon > 0$ there exist $N, M \in \mathbb{N}$ such that $|\mathcal{L}_{q,M,w}(U) - \mathrm{ED}_q(p_{\text{data}}, U)| < \varepsilon$ a.s..*

Proof. For N data points $\mathbf{x}_+^i \sim p_{\text{data}}$ and perturbed points $\mathbf{y}^i \sim q(\cdot|\mathbf{x}_+^i)$ denote the M corresponding negative samples by $\mathbf{x}_-^{i,j} \sim p_{\text{neg},q,\mathbf{y}^i}$. Notice that the distribution of the negative samples depends on $\mathbf{y}^i$. Using the triangle inequality, we can upper bound the difference $|\mathrm{ED}_q(p_{\text{data}}, U) - \mathcal{L}_{q,M,w}(U)|$ by upper bounding the following two terms, individually:

$$\begin{aligned} &\left| \mathrm{ED}_q(p_{\text{data}}, U) - \frac{1}{N} \sum_{i=1}^{N} \log \mathbb{E} \left[\exp(U(\mathbf{x}_+^i) - U(\mathbf{x}_-^{i,j}) \,\middle|\, \mathbf{x}_+^i, \mathbf{y}^i \right] \right| \\ &\qquad + \left| \frac{1}{N} \sum_{i=1}^{N} \log \mathbb{E} \left[\exp(U(\mathbf{x}_+^i) - U(\mathbf{x}_-^{i,j}) \,\middle|\, \mathbf{x}_+^i, \mathbf{y}^i \right] - \mathcal{L}_{q,M,w}(U) \right| \end{aligned}$$

The conditioning expresses that the expectation is only taken in $\mathbf{x}_-^{i,j} \sim p_{\text{neg},q,\mathbf{y}^i}$ while keeping the values of the random variables $\mathbf{x}_+^i$ and $\mathbf{y}^i$ fixed. The first term can be bounded by a sequence $\varepsilon_N \xrightarrow{a.s.} 0$ due to the normal strong law of large numbers. For the second term one needs to consider that the distribution $p_{\text{neg},q,\mathbf{y}^i}$ depends on the random variable $\mathbf{y}^i$. For this reason, we notice that $\mathbf{x}_-^{i,j}$

are conditionally indepedent given $\mathbf{x}_+^i, \mathbf{y}^i$ and employ a conditional version of the strong law of large numbers (Majerek et al., 2005, Theorem 4.2) to obtain

$$\frac{1}{M}\sum_{j=1}^{M}\exp\left(U(\mathbf{x}_+^i) - U(\mathbf{x}_-^{i,j})\right) \xrightarrow{a.s.} \mathbb{E}\left[\exp(U(\mathbf{x}_+^i) - U(\mathbf{x}_-^{i,j}) \,\middle|\, \mathbf{x}_+^i, \mathbf{y}^i\right]$$

Next, we have that the deterministic sequence $w/M \to 0$. Thus, adding the stabilisation w/M does not change the limit in M. Furthermore, since the logarithm is continuous, the limit also holds after applying the logarithm. Finally, the estimate translates to the sum by another application of the triangle inequality: For each $i = 1, 2, \ldots, N$ there exists a sequence $\varepsilon_{i,M} \xrightarrow{a.s.} 0$ such that

$$\begin{aligned}
&\left|\frac{1}{N}\sum_{i=1}^{N}\log\mathbb{E}\left[\exp(U(\mathbf{x}_+^i) - U(\mathbf{x}_-^{i,j}) \,\middle|\, \mathbf{x}_+^i, \mathbf{y}^i\right] - \mathcal{L}_{q,M,w}(U)\right| \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\left|\log\mathbb{E}\left[\exp(U(\mathbf{x}_+^i) - U(\mathbf{x}_-^{i,j}) \,\middle|\, \mathbf{x}_+^i, \mathbf{y}^i\right] - \log\frac{1}{M}\sum_{j=1}^{M}\exp\left(U(\mathbf{x}_+^i) - U(\mathbf{x}_-^{i,j})\right)\right| \\
&< \frac{1}{N}\sum_{i=1}^{N}\varepsilon_{i,M} \leq \max(\varepsilon_{1,M}, \ldots, \varepsilon_{N,M})\,.
\end{aligned}$$

Hence, for each $\varepsilon > 0$ there exists an $N \in \mathbb{N}$ and an $M(N) \in \mathbb{N}$ such that $|\mathrm{ED}_q(p_{\mathrm{data}}, U) - \mathcal{L}_{q,M(N),w}(U)| < \varepsilon$ almost surely. □

D Related Work

Contrastive loss functions Our work is based on an unpublished work on energy discrepancies in the continuous case (Schröder et al., 2023). The motivation for such constructed loss functions lies in the data processing inequality. A similar loss has been suggested before as KL contraction divergence (Lyu, 2011), however, only for its theoretical properties. Interestingly, the structure of the stabilised energy discrepancy loss shares similarities with other contrastive losses such as Ceylan & Gutmann (2018); Gutmann & Hyvärinen (2010); van den Oord et al. (2018). This poses the question of possible classification-based interpretations of energy discrepancy and of the w-stabilisation.

Contrastive divergence and Sampling. Discrete training methods for energy-based models largely rely on contrastive divergence methods, thus motivating a lot of work on discrete sampling and proposal methods. Improvements of the standard Gibbs method were proposed by Zanella (2020) through locally informed proposals. The method was extended to include gradient information (Grathwohl et al., 2021) to drastically reduce the computational complexity of flipping bits of binary valued data and to flipping bits in several places (Sun et al., 2022b; Emami et al., 2023; Sun et al., 2022a). Finally, discrete versions of Langevin sampling have been introduced based on this idea (Zhang et al., 2022b; Rhodes & Gutmann, 2022; Sun et al., 2023). Consequently, most current implementations of contrastive divergence use multiple steps of a gradient based discrete sampler. Alternatively, energy-based models can be trained using generative flow networks which learns a Markov chain to construct data by optimising a given reward function. The Markov chain can be used to obtain samples for contrastive divergence without MCMC from the EBM (Zhang et al., 2022a).

Other training methods for discrete EBMs. There also exist some MCMC free approaches for training discrete EBMs. Our work is most similar to concrete score matching (Meng et al., 2022) which uses neighbourhood structures to define a replacement of the continuous score function. Another sampling free approach for training discrete EBMs is ratio matching (Hyvärinen, 2007; Lyu, 2012). However is has been found that also for ratio matching, gradient information drastically improves the performance (Liu et al., 2023). Moreover, Dai et al. (2020) proposed to apply variational approaches to train discrete EBMs instead of MCMC. Eikema et al. (2022) replaced 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).

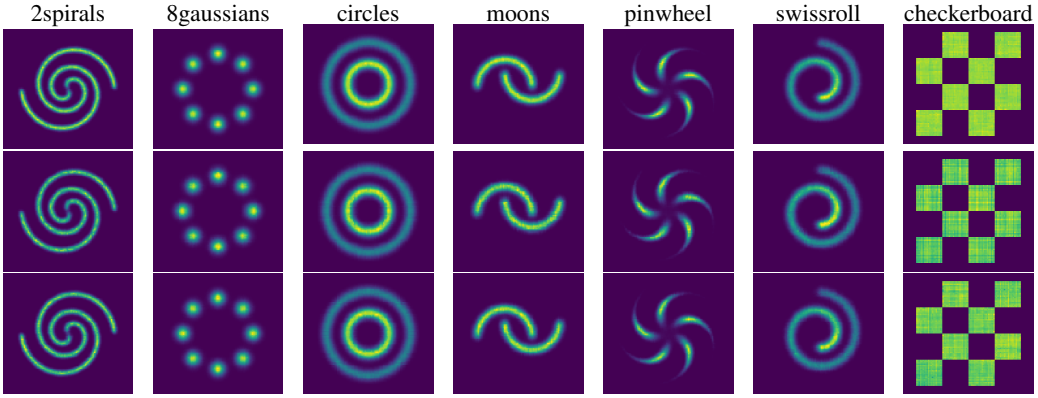

Figure 3: Visualization of the energy function. Top to bottom: ED-Bern, ED-Pool, ED-Grid.

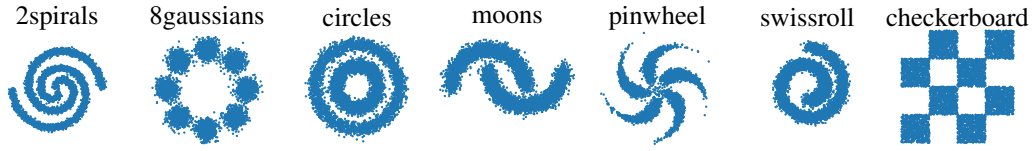

Figure 4: Visualization of samples for discrete density estimation from ground truth.

E More about Experiments

E.1 Training Ising Models

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 ED-Bern, and the window side is $\sqrt{D} \times \sqrt{D}$ in ED-Pool. The parameter J_ϕ is learned by the Adam (Kingma & Ba, 2014) optimizer 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 regularization with a coefficient in $\{10, 5, 1, 0.1, 0.01\}$ 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.

Table 3: 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).

	$D = 10^2$					$D = 9^2$	
Method \ σ	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 (ours)	5.1	4.0	2.9	2.5	**2.3**	5.1	4.3
ED-Pool (ours)	4.9	3.6	3.2	**2.6**	**2.3**	4.9	3.6
ED-Grid (ours)	4.6	4.0	3.1	**2.6**	**2.3**	4.5	4.0

Quantitative Results. 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 ture matrix J. As shown in Table 3, 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.

E.2 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

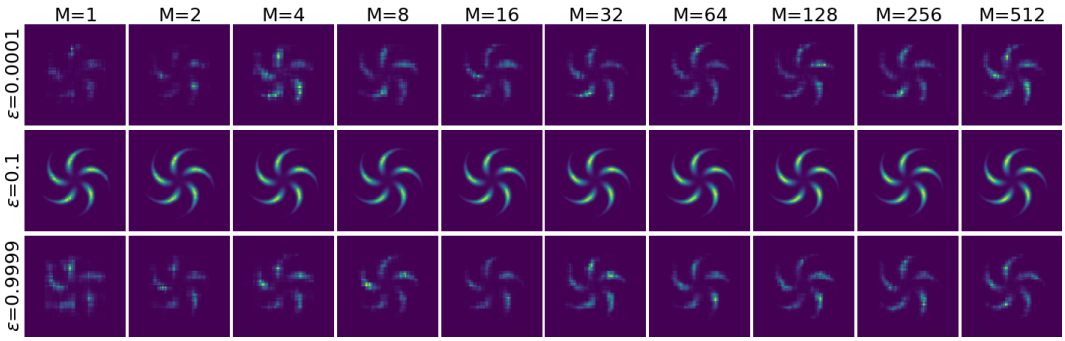

Figure 5: Density estimation results of ED-Bern on the pinwheel with different ϵ, M and $w = 1$.

to obtain a 32-bits vector $\mathbf{x}$. As a result, the probabilistic mass function in the discrete space is $p(\mathbf{x}) \propto \hat{p}\left([\text{GrayToFloat}(\mathbf{x}_{1:16}), \text{GrayToFloat}(\mathbf{x}_{17:32})]\right)$. It is noteworthy that learning on this discrete space presents challenges due to the highly non-linear nature of the Gray code transformation.

The energy function is parameterised by a 4 layer MLP with 256 hidden dimensions and Swish (Ramachandran et al., 2017) activation. We train the EBM for 10^5 steps and adopt an Adam optimiser with a learning rate of 0.002 and a batch size of 128 to update the parameter. For the energy discrepancy, we choose $w = 1, M = 32$ for all variants, $\epsilon = 0.1$ in ED-Bern, and the window size is 32×1 in ED-Pool. 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. We qualitatively visualise the learned energy functions of our proposed approaches in Figure 3. To provide further insights into the oracle energy landscape, we also plot the ground truth samples in Figure 4. The results clearly demonstrate that energy discrepancy effectively fits the data distribution, validating the efficacy of our methods.

The Effect of ϵ in Bernoulli Perturbation. Perhaps surprisingly, we find that the proposed energy discrepancy loss with Bernoulli perturbation is very robust to the noise scalar ϵ. In Figure 6, w visualise the learned energy landscapes with different ϵ. The results demonstrate that ED-Bern is able to learn faithful energy functions, even with extreme values of ϵ, such as $\epsilon \in \{0.999, 0.001\}$. This highlights the robustness and effectiveness of our approach. In Figure 5, we further show that, with $\epsilon \in \{0.9999, 0.0001\}$, ED-Bern can still learn a faithful energy landscape using a large value of M. However, when $\epsilon \in \{1, 0\}$, ED-Bern fails to work. It is noteworthy that the choice of ϵ is highly dependent on the specific structure of the dataset. While ED-Bern exhibits robustness to different values of ϵ in the synthetic data, we have observed that a large value of ϵ ($\epsilon \geq 0.1$) is not effective for discrete image modeling.

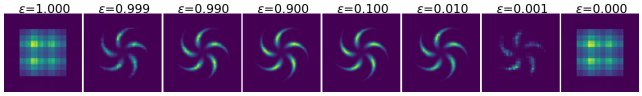

Figure 6: Density estimation results of ED-Bern on the pinwheel with different ϵ and $M = 32, w = 1$.

The Effect of Window Size in Deterministic Transformation. To investigate the effectiveness of the window size in ED-Pool, we conduct experiments in Figure 7 with different window sizes. The results indicate that employing a small window size (*e.g.*, 2×1) does not provide sufficient information for energy discrepancy to effectively learn the underlying data structure. Furthermore, our empirical findings suggest that solely increasing the value of M is not a viable solution to address this issue. Again, the choice of the window size should depend on the underlying data structure. In the discrete image modelling, we find that even with a small window size

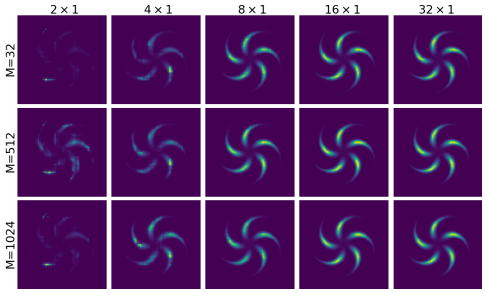

Figure 7: Density estimation results of ED-Pool on the pinwheel with different window sizes, M and $w = 1$.

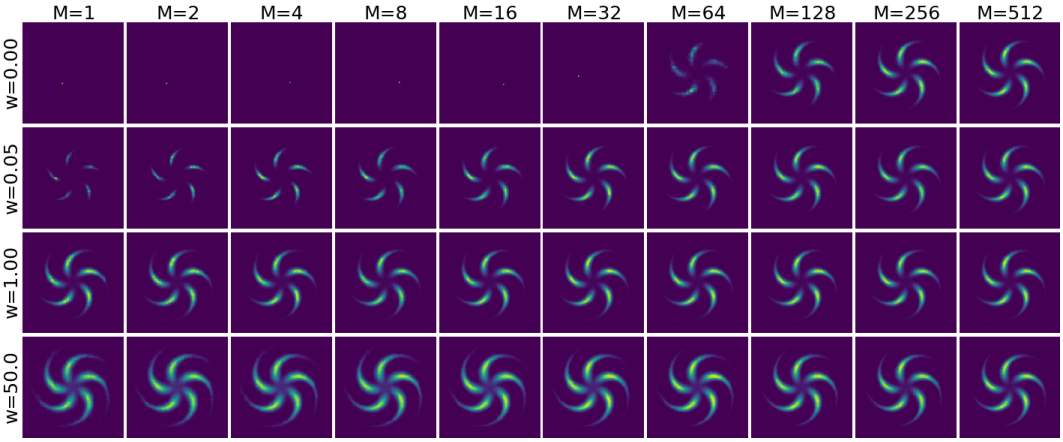

Figure 8: Density estimation results of ED-Bern on the pinwheel with different w, M and $\epsilon = 0.1$.

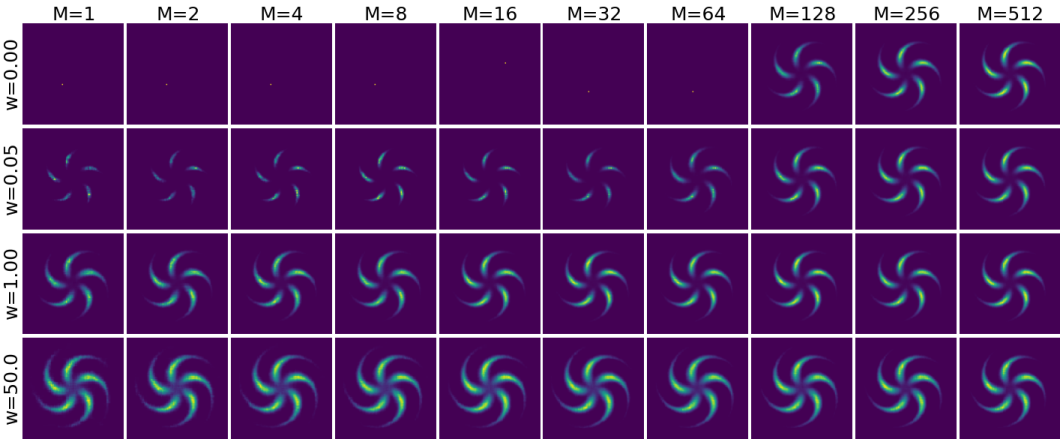

Figure 9: Density estimation results of ED-Pool on the pinwheel with different w, M and the window size is 32×1.

(*i.e.*, 4×4), energy discrepancy yields an energy with low values on the data-support but rapidly diverging values outside of it. Therefore, it fails to learn a faithful energy landscape.

Qualitatively Understanding the Effect of w and M. The hyperparameters w and M play a crucial role in the estimation of energy discrepancy. Increasing M can reduce the variance of the Monte Carlo estimation of the contrastive potential in (1), while a proper value of w can improve the stabilisation of training. Here, we evaluate the effect of w and M on the variants of energy discrepancy in Figures 8 to 10. Based on empirical observations, we observe that when $w = 0$ and M is small (*e.g.*, $M \leq 32$ for ED-Bern and $M \leq 64$ for ED-Pool and ED-Grid), energy discrepancy demonstrates rapid divergence and fails to converge. Additionally, we find that increasing M can address this issue to some extent and introducing a non-zero value for w can significantly stabilize the convergence, even with $M = 1$. Moreover, larger w tends to produce a flatter estimated energy landscapes, which also aligns with the findings in continuous scenarios of energy discrepancy Schröder et al. (2023).

E.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, $\epsilon = 0.001$ for ED-Bern, and the window size is 2×2 for ED-Pool. Note that here we choose a relatively small ϵ and window size, since we empirically find that the loss of energy discrepancy converges to a constant rapidly with larger ϵ and window size, which can not

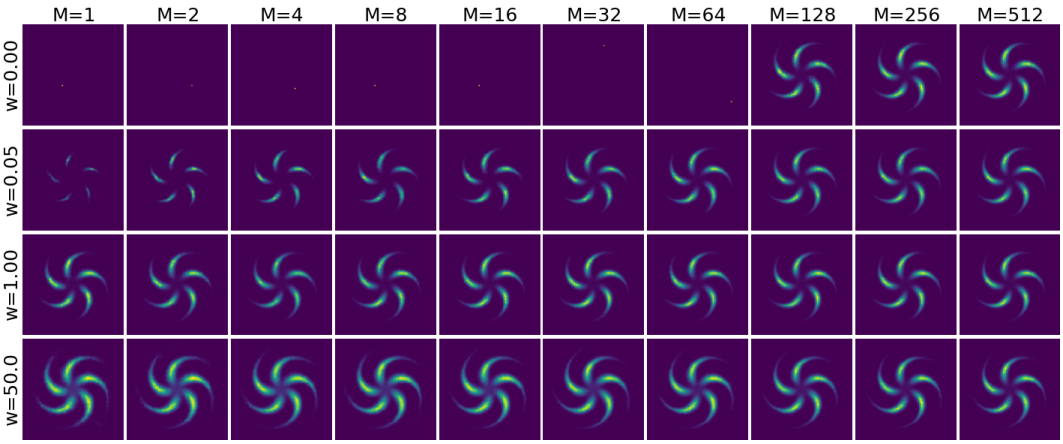

Figure 10: Density estimation results of ED-Grid on the pinwheel with different w, M.

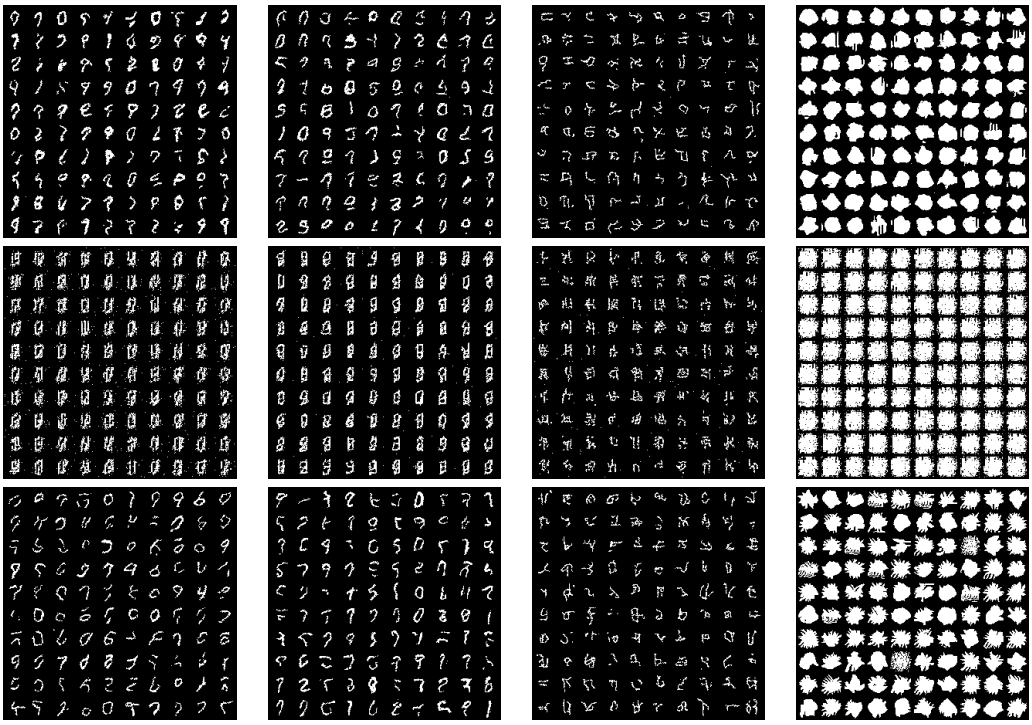

Figure 11: Generated samples on discrete image modelling. Left to right: Static MNIST, Dynamic MNIST, Omniglot, Caltech Silhouettes. Top to bottom: ED-Bern, ED-Pool, ED-Grid.

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 $5,000$ iterations by conducting Annealed Importance Sampling (AIS) with a discrete Langevin 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. We show the generated images in Figure 11, which are the samples in the final step of AIS. We see that our methods can generate realistic images on the Omniglot dataset but mediocre images on Caltech Silhouette. We hypothesise that improving the design of the affinity structure in the neighborhood-based transformation can lead to better results. On both the static and dynamic MNIST datasets, ED-Bern and ED-Grid generate diverse and high-quality images. However, ED-Pool experiences mode collapse, resulting in limited variation in the generated samples.