Energy Discrepancies

Andrew Duncan
Imperial College London

Joint work with:

Tobias Schröder
Imperial College London

Zijing Ou
Imperial College London

Jen Ning Lim
University of Warwick

Yingzhen Li
Imperial College London

Sebastian F. Vollmer
DFKI and RPTU
Kaiserslautern

Unnormalized Statistical Models

Models with an intractable normalisation constant in the likelihood

$$p_\theta(x) = \frac{e^{-U_\theta(x)}}{Z_\theta},$$

where the normalisation constant Z_θ which is expensive to compute.

In many applications, p_θ belong to some exponential family, i.e.

$$U_\theta(x) = \theta^\top T(x),$$

where $\theta = (\theta_1, \ldots, \theta_p)$.

Examples

Ising Model

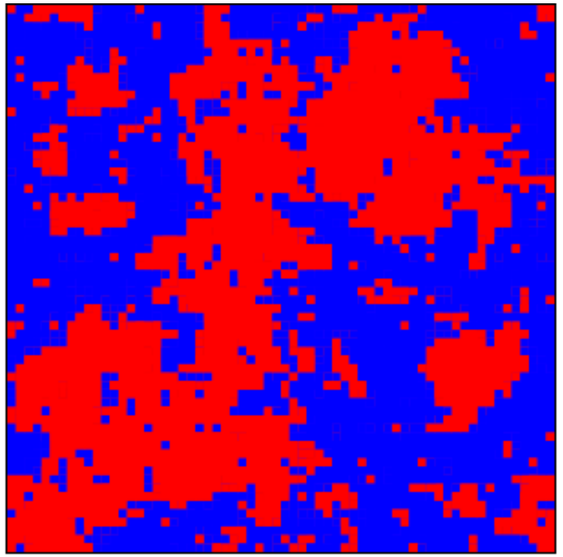
Let $y = (y_1, \ldots, y_n)$ where the y_i's are binary variables lying on a lattice.

We consider the probability density

$$p_\theta(y) = \frac{\exp\left(\alpha \sum_i y_i + \beta \sum_{i \sim j} \mathbf{1}[y_i = y_j\right)}{Z_\theta}$$

where $\theta = (a, b)$.

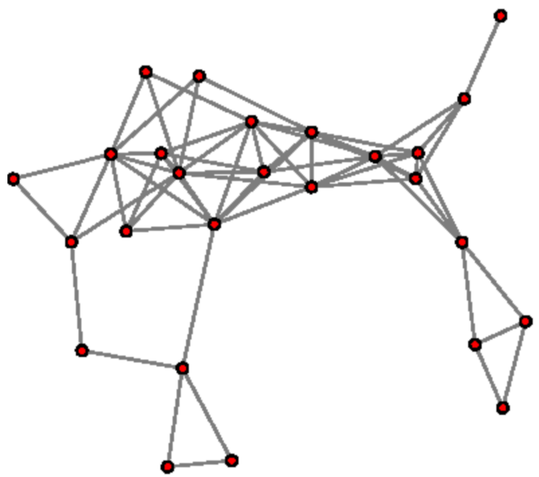
- The normalisation constant is a sum of 2^n terms.



Examples

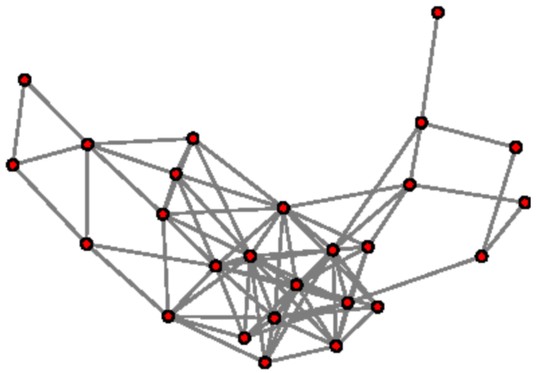
Exponential Random Graphs

Consider a random graph with edge locations labelled $1, \ldots, p$.



A realisation of the graph is defined by $y = (y_1, \ldots, y_p)$ where

- Edge i is present if $y_i = 1$;
- Absent if $y_i = 0$.



Exponential graph model:

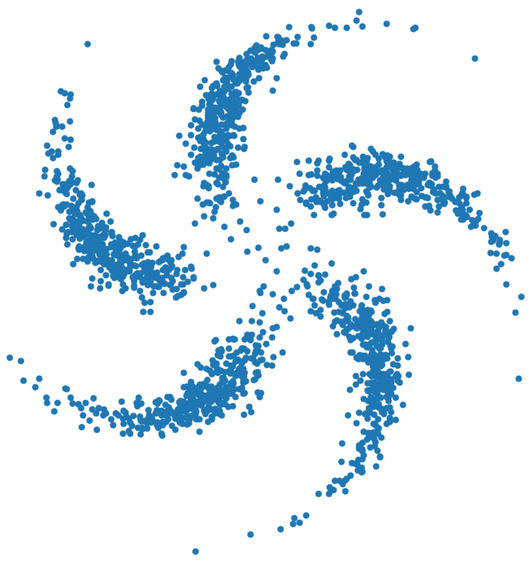
$$p_\theta(y) = \frac{\exp(\theta^\top s(y))}{Z_\theta}.$$

Examples

Energy Function (Neural Network)

Energy-Based Models (LeCun, 2006)

Data

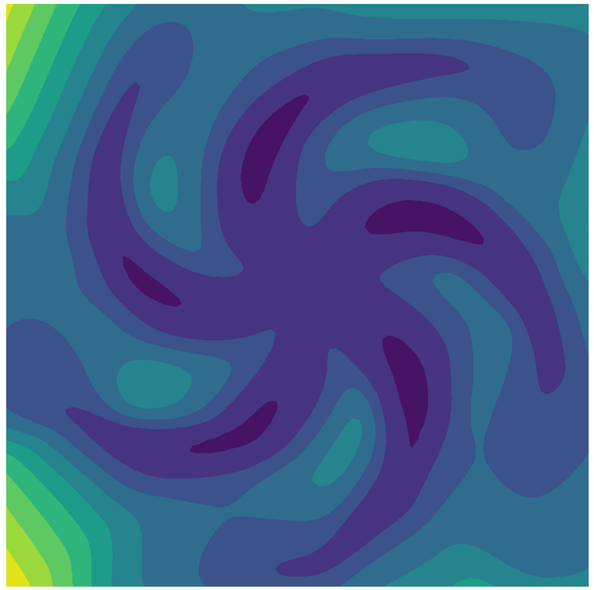


$$\mathbf{x}^i \overset{i.i.d.}{\sim} p_{\text{data}}$$

Model distribution

$$p_\theta(\mathbf{x}) = \frac{\exp(-U_\theta(\mathbf{x}))}{Z_\theta}$$

Learned Energy Function



$$U_\theta(\mathbf{x})$$

Intractable normalisation (partition function): $Z_\theta = \int \exp(-U_\theta(\mathbf{x}))\mathrm{d}\mathbf{x}$

EBMs: Likelihood-free Inference

Glaser et al. (2023)

- Noisy simulator of data with unknown likelihood:

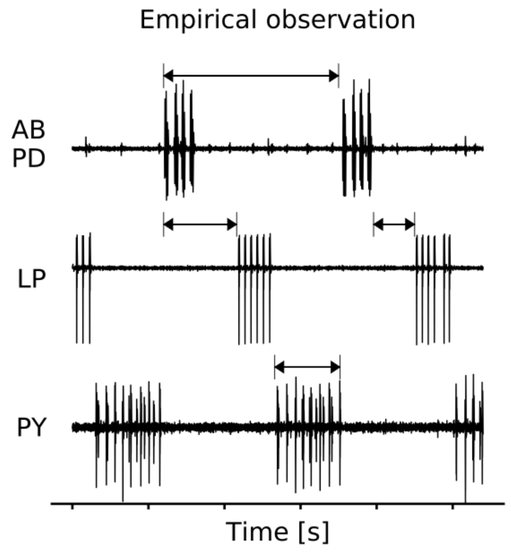
$$\mathbf{x} = G(\boldsymbol{\phi}, \boldsymbol{\epsilon})$$

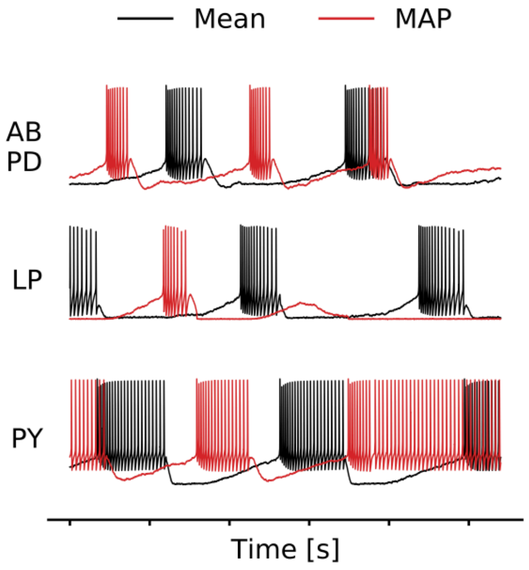
- Learn joint distribution:

$$p_\theta(\mathbf{x}, \boldsymbol{\phi}) = \frac{\exp(-U_\theta(\mathbf{x}, \boldsymbol{\phi}))p(\boldsymbol{\phi})}{Z_\theta}$$

- Infer posterior over parameters for observations $\mathbf{x}^o$:

$$p_\theta(\boldsymbol{\phi} \,|\, \mathbf{x}^o) = \frac{\exp(-U_\theta(\mathbf{x}^o, \boldsymbol{\phi}))p(\boldsymbol{\phi})}{Z_{\theta,\mathbf{x}^o}}$$





Glaser et al. (2023)
Maximum Likelihood Learning of Unnormalized Models for Simulation-Based Inference

EBMs & Classifiers

Grathwohl et al. (2020)

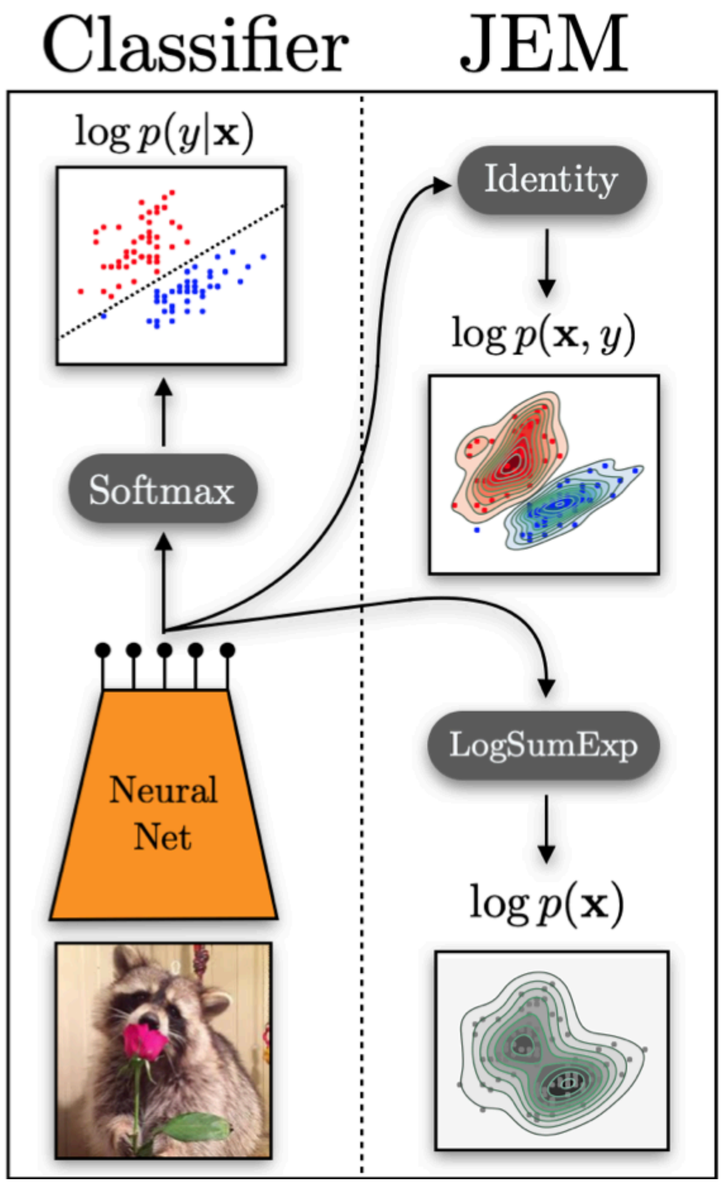
Train EBM on tuples $(\mathbf{x}, \mathbf{y})$

$$p_\theta(\mathbf{x}, \mathbf{y}) = \frac{\exp(-U_\theta(\mathbf{x}, \mathbf{y}))}{\int \exp(-U_\theta(\mathbf{x}', \mathbf{y}'))\mathrm{d}\mathbf{x}'\mathrm{d}\mathbf{y}'}$$

Data points Labels

Classifier: $p(\mathbf{y} \mid \mathbf{x}) = \mathrm{softmax}_{\mathbf{y}}(-U_\theta(\mathbf{x}, \mathbf{y}))$

Unconditional EBM: $U_\theta(\mathbf{x}) = -\,\mathrm{LSE}_{\mathbf{y}}(-U_\theta(\mathbf{x}, \mathbf{y}))$



Grathwohl et al. (2020)
Your classifier is secretly an energy-based model and you should treat it like one

EBMs & Plugin Priors

- Noisy forward process

$$\mathbf{z} = T(\mathbf{x}) + \boldsymbol{\epsilon}$$

- Recovery via Maximum a posteriori estimate

$$\hat{\mathbf{x}} = \operatorname{argmax}_{\mathbf{x}} p(\mathbf{x} \,|\, \mathbf{z}^o)$$

$$= \operatorname{argmin}_{\mathbf{x}} \frac{1}{2\lambda} \|\mathbf{z}^o - T(\mathbf{x})\|^2 \underbrace{-\log p_{\text{data}}(\mathbf{x})}_{\approx U_\theta(\mathbf{x})}$$

Infer online

Learn offline

Inference for Unnormalised Models

Suppose we observe samples $y_1, \ldots, y_n$.

Objective: Infer $\theta \in \Theta$ such that

$$p_\theta(x) = \frac{\exp(-E_\theta(x)}{Z_\theta}$$

best explains the explains.

- Challenge arises from dependence of Z_θ on θ.
- In the Bayesian setting, gives rise to *doubly-intractable* problems.

Inference for Unnormalised Models

MCMC-MLE (Geyer, 1994)

Log-Likelihood:

$$l_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(y_i) - \log Z(\theta).$$

Sample using MCMC

Idea: Estimate θ by maximising the function

$$\hat{l}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p_\theta(y_i)}{p_\psi(y_i)} - \log\left(\frac{1}{m}\sum_{j=1}^{m} \frac{p_\theta(x_j)}{p_\psi(x_j)}\right)$$

where the x_j's are m artificial data-points from a user-chosen distribution p_ψ .

Approximation of the log-likelihood ratio $l_n(\theta) - l_n(\psi)$.

Inference for Unnormalised Models

Noise Contrastive Estimation (Gutmann & Hyvarinen, 2012)

Maximise likelihood of logistic classifier

$$l_{n,m}^{NCE}(\theta, \nu) = \sum_{i=1}^{n} \log q_{\theta,\nu}(y_i) + \sum_{i=1}^{m} \log(1 - q_{\theta,\nu}(x_i)).$$

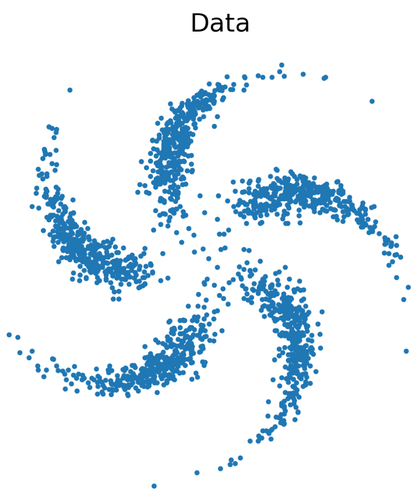
where $q_{\theta,\nu}(x)$ is the probability of a label 1 for a value x and

$$\log\left(\frac{q_{\theta,\nu}(x)}{1 - q_{\theta,\nu}(x)}\right) = \log\left(\frac{p_\theta(x)}{p_\psi(x)}\right) + \nu + \log(m/n)\,.$$

Again the x_j's are m artificial data-points from a user-chosen distribution p_ψ .

Inference for Unnormalised Models

Contrastive Divergence (Hinton 1999)



Taking gradient of log likelihood:

$$-\partial_\theta l_n(\theta) = \partial_\theta \log Z(\theta) - \frac{1}{n}\sum_{i=1}^{n} \partial_\theta \log p_\theta(y_i)$$

$$\approx \frac{1}{m}\sum_{j=1}^{m} \partial_\theta \log p_\theta(x_j) - \frac{1}{n}\sum_{i=1}^{n} \partial_\theta \log p_\theta(y_i)\,.$$

where the x_j's are sampled using MCMC targeting p_θ.

Inference for Unnormalised Models

Training Products of Experts by Minimizing Contrastive Divergence

GCNU TR 2000-004

Geoffrey E. Hinton
Gatsby Computational Neuroscience Unit
University College London
17 Queen Square, London WC1N 3AR, U.K.

http://www.gatsby.ucl.ac.uk/

Inference for Unnormalised Models

Training Products of Experts by Minimizing Contrastive Divergence

GCNU TR 2000-004

The intuitive motivation for using this "contrastive divergence" is that we would like the Markov chain that is implemented by Gibbs sampling to leave the initial, distribution Q^0 over the visible variables unaltered. Instead of running the chain to equilibrium and comparing the initial and final derivatives we can simply run the chain for one full step and then update the parameters to reduce the tendency of the chain to wander away from the initial distribution on the first step. Because Q^1 is one step closer to the equilibrium distribution than Q^0, we are guaranteed that $Q^0||Q^\infty$ exceeds $Q^1||Q^\infty$ unless Q^0 equals Q^1, so the contrastive divergence can never be negative. Also, for Markov chains in which all transitions have non-zero probability, $Q^0 = Q^1$ implies $Q^0 = Q^\infty$ so the contrastive divergence can only be zero if the model is perfect.

Inference for Unnormalised Models

The divergence yielding CD gradient descent?

$$\textbf{CD}(p_{data}\,||\,p_\theta) = \textbf{KL}(p_{data}\,||\,p_\theta) - \textbf{KL}(\Pi_\theta^{(t)} p_{data}\,||\,p_\theta)$$

where $\Pi_\theta^{(t)} p_{data}$ is the probability distribution p_{data} after t-steps of a Markov chain with stationary distribution p_θ starting from p_{data}.

- Data Processing Inequality $\Rightarrow$ CD$(p_{data}\,||\,p_\theta) \geq 0$.
- CD$(p_{data}\,||\,p_\theta) = 0 \Rightarrow p_\theta = p_{data}$.

Inference for Unnormalised Models

The divergence yielding CD gradient descent?

Taking gradients:

$$\partial_\theta D(p_{data} \,||\, p_\theta) = \mathbb{E}_{p^{(t)}_{data,\theta}} \partial_\theta \log p_\theta - \mathbb{E}_{p_{data}} \partial_\theta \log p_\theta$$

$$+ \partial_\theta \Pi^{(t)}_\theta p_{data} \nabla_q \mathbf{KL}(q \,|\, p_\theta) \Big|_{q = \Pi^{(t)}_\theta p_{data}}$$

Inference for Unnormalised Models

The divergence yielding CD gradient descent?

Taking gradients:

$$\partial_\theta D(p_{data} \,||\, p_\theta) = \mathbb{E}_{p^{(t)}_{data,\theta}} \partial_\theta \log p_\theta - \mathbb{E}_{p_{data}} \partial_\theta \log p_\theta$$

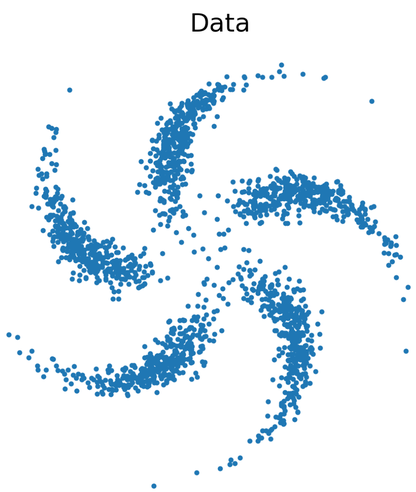
$$\cancel{+ \partial_\theta \Pi^{(t)}_\theta p_{data} \nabla_q \mathbf{KL}(q \,|\, p_\theta) \Big|_{q = \Pi^{(t)}_\theta p_{data}}}$$

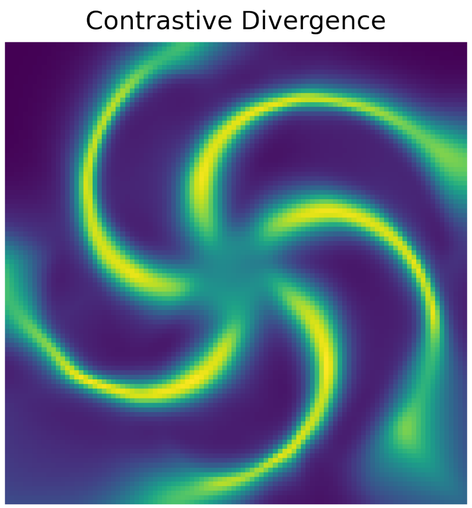
Inference for Unnormalised Models

Contrastive Divergence:

CD is zero when $p_\theta = p_{data}$, BUT:

- Is not the gradient of any objective function.
- Dropping that term can drastically affect stability of the optimisation process.
- Tends to produce “malformed energies”
- There have been attempts to correct this using entropy estimators *[Du, Li, Tenenbaum, Mordatch, 2021]*





Training Energy-Based Models is hard

Score Matching

- Minimises distance between (Stein) scores of data and model:

$$\mathbb{E}_{p_{\text{data}}(\mathbf{x})}\left[\|\nabla \log p_{\text{data}}(\mathbf{x}) - \nabla \log p_\theta(\mathbf{x})\|^2\right]$$

- Via an integration of parts one obtains:

$$\text{SM}(p_{\text{data}}, p_\theta) = \mathbb{E}_{p_{\text{data}}(\mathbf{x})}\left[-\Delta_{\mathbf{x}} U_\theta(\mathbf{x}) + \frac{1}{2}\|\nabla_{\mathbf{x}} U_\theta(\mathbf{x})\|^2\right]$$

Training Energy-Based Models is hard

Score Matching

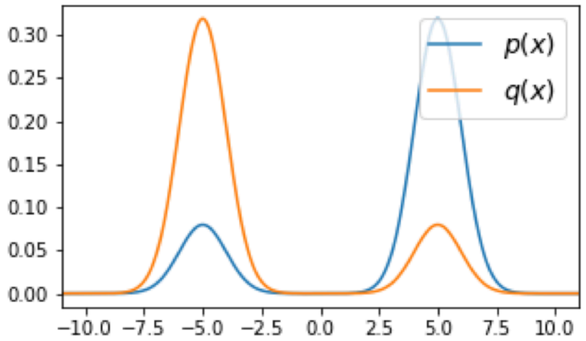
- Minimises distance between (Stein) scores of data and model:

$$\mathbb{E}_{p_{\text{data}}(\mathbf{x})}\left[\|\nabla \log p_{\text{data}}(\mathbf{x}) - \nabla \log p_\theta(\mathbf{x})\|^2\right]$$

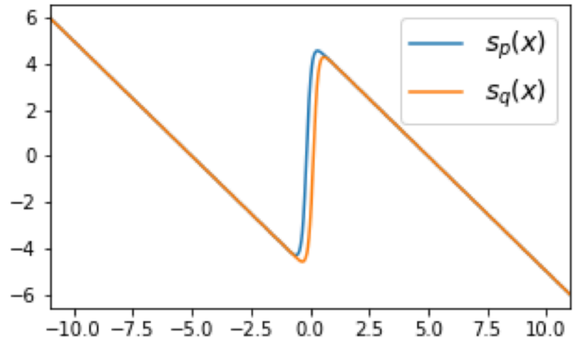
- Via an integration of parts one obtains:

$$\text{SM}(p_{\text{data}}, p_\theta) = \mathbb{E}_{p_{\text{data}}(\mathbf{x})}\left[\Delta_{\mathbf{x}} \log p_\theta(\mathbf{x}) + \frac{1}{2}\|\nabla_{\mathbf{x}} \log p_\theta(\mathbf{x})\|^2\right]$$

- Blind to mixture weights (Nearsightedness)



(a) Densities of p and q



(b) Score functions of p and q

Zhang et al. (2022)
Towards Healing the Blindness of Score Matching

Training Energy-Based Models is hard

Diffusion Score Matching (Barp, Briol et al, 2022).

- More generally, we can define the diffusion score matching discrepancy

$$D(p_{data}, p) = \int \frac{1}{2} \| b(x)^\top \nabla \log p(x) \|^2 + \nabla \cdot (\Sigma(x) \nabla \log p(x))\, p_{data}(dx),$$

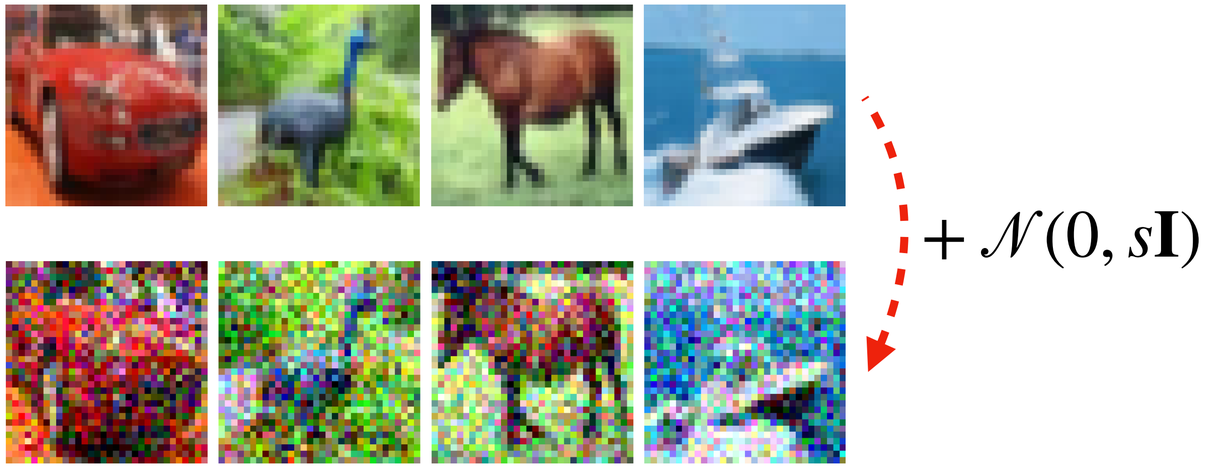
where $\Sigma(x) = b(x) b(x)^\top$, is positive definite.

Energy Discrepancy: Motivation

Spread out distributions with Gaussian noise to overcome mode blindness

$$p_s(\mathbf{y}) := \int \gamma_s(\mathbf{y} - \mathbf{x}) p_{\text{data}}(\mathbf{x}) \mathrm{d}\mathbf{x}$$

$$\exp(-U_s(\mathbf{y})) := \int \gamma_s(\mathbf{y} - \mathbf{x}) \exp(-U(\mathbf{x})) \mathrm{d}\mathbf{x}$$

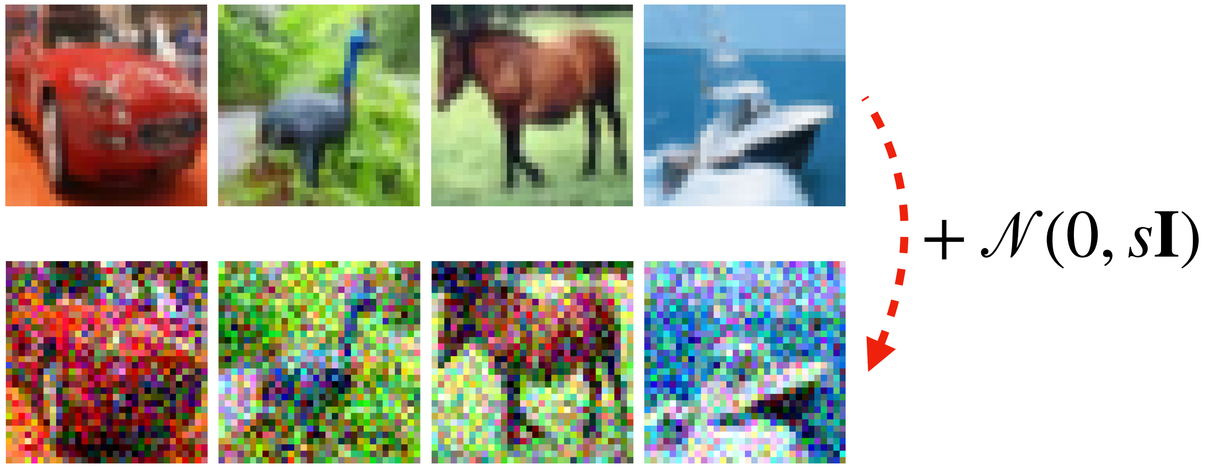


Energy Discrepancy: Motivation

Spread out distributions with Gaussian noise to overcome mode blindness

$$p_s(\mathbf{y}) := \int \gamma_s(\mathbf{y} - \mathbf{x}) p_{\text{data}}(\mathbf{x}) \mathrm{d}\mathbf{x}$$

$$\exp(-U_s(\mathbf{y})) := \int \gamma_s(\mathbf{y} - \mathbf{x}) \exp(-U(\mathbf{x})) \mathrm{d}\mathbf{x}$$



$$\mathsf{ED}_{\gamma_s}(p_{data}, p) = \underbrace{\mathbb{E}_{p_{\text{data}}(\mathbf{x})}[U(\mathbf{x})] - \mathbb{E}_{p_s(\mathbf{y})}[U_s(\mathbf{y})]}_{\text{Gaussian Energy Discrepancy}}$$

Energy Discrepancy

Energy-Based distribution: $p(\mathbf{x}) \propto \exp(-U(\mathbf{x}))$

Conditional (noising) distribution: $q(\mathbf{y} \mid \mathbf{x})$

Contrastive potential: $U_q(\mathbf{y}) = -\log \int \exp(-U(\mathbf{x}))q(\mathbf{y} \mid \mathbf{x})\mathrm{d}\mathbf{x}$

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

Energy Discrepancy

Energy-Based distribution: $p(\mathbf{x}) \propto \exp(-U(\mathbf{x}))$

Conditional (noising) distribution: $q(\mathbf{y} \mid \mathbf{x})$

Contrastive potential: $U_q(\mathbf{y}) = -\log \int \exp(-U(\mathbf{x}))q(\mathbf{y} \mid \mathbf{x})\mathrm{d}\mathbf{x}$

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

- Data processing inequality implies that $\mathrm{ED}_q(p_{data}, p) \geq 0$.
- Energy Discrepancy is functionally convex in U.
- For nice q, ED has a unique global minimiser at $\exp(-U^*) \propto p_{\mathrm{data}}$

Energy Discrepancy: Connection to CD

Back to the original CD Loss

$$D(p_{data}||p_\theta) = \mathrm{KL}(p_{data}||p_\theta) - \mathrm{KL}(p^{(t)}_{data,\theta}||p_\theta)$$

$$= \mathrm{KL}(p_{data}||p_\theta) - \mathrm{KL}(\Pi^{(t)}_\theta p_{data}||\Pi^{(t)}_\theta p_\theta)$$

since p_θ is an invariant wrt $\Pi^{(t)}_\theta$.

Idea: Replace $\Pi^{(t)}_\theta$ with something not depending on θ.

Energy Discrepancy: Limiting Behaviour

Let q_t be the transition density of the diffusion process

$$dX_t = a(X_t)\,dt + b(X_t)dW_t,$$

And let $\Sigma(x) = b(x)b(x)^\top$

Given:

Diffusion Score Matching Discrepancy

- $p(x) \propto e^{U(x)}$
- $U_{q_t}(\,\cdot\,) = -\log \int q_t(\,\cdot\,|x)e^{U(x)}\,dx$ and $p_t(\,\cdot\,) = \int q_t(\,\cdot\,|x)p_{data}(dx)$.

Then we have a de Bruijin type relationship:

$$ED_{q_t}(p_{data}, p) = \int_0^t \mathbb{E}_{p_s(x_s)}\left[-\sum_{i,j=1}^{d} \partial_{x_j}(\Sigma_{ij}(x)\partial_{x_i}U_{q_s}(x_s)) + \frac{1}{2}\|b^\top(x_s)\,\nabla U_{q_s}(x_s)\|^2\right] ds + \text{const.}$$

Energy Discrepancy: Limiting Behaviour

In the case when the drift is zero (i.e. Gaussian Perturbation).

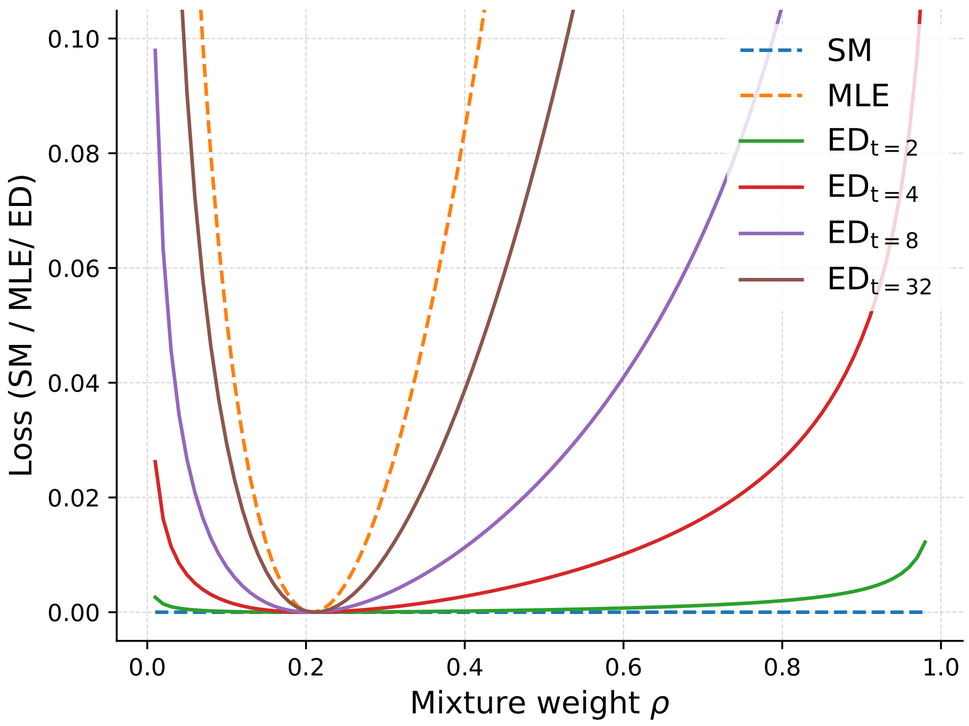
Then

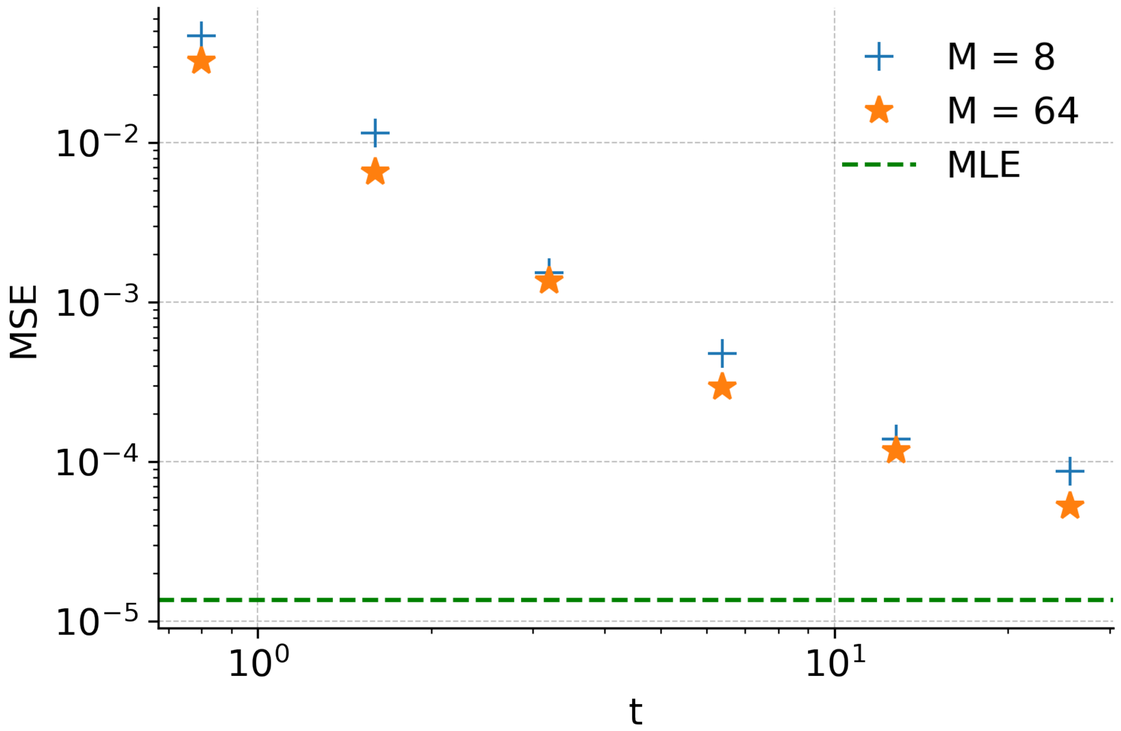
$$\left| \mathsf{ED}_{q_t}(p_{data}, p) + \mathbb{E}_{p_{data}(x)}[\log p(x)] - c(t) \right| \leq \frac{1}{2t} W_2^2(p_{data}, p),$$

where $c(t)$ is a constant independent of p.

- In particular, as $t \to \infty$, ED behaves like maximum likelihood estimation.
- Can be generalised to OU process perturbation $dX_t = \alpha X_t \, dt + dW_t$, provided $\alpha < 0$.

Choice of Perturbation





Backward Approximation

Let q_t be the transition density of the Ito diffusion

$$dX_t = a(X_t)\,dt + b(X_t)dW_t.$$

Then by Feynman-Kac formula we have

$$U_t(y) = -\log \mathbb{E}_y \left[\exp\left(\int_0^t c(y_s)\,ds - U(y_t) \right) \right],$$

where y_t is the backwards (time-reversed) SDE.

This yields an approximation

$$U_t(\mathbf{x}_t) \approx -\log \frac{1}{M} \sum_{j=1}^{M} \exp\left(\left(\sum_{k=1}^{K} c(\tilde{\mathbf{y}}_{t_k}^j)(t_k - t_{k-1}) \right) - U(\tilde{\mathbf{y}}_t^j) \right)$$

Forward Approximation

$$\mathbb{E}_{p_{\text{data}}(\mathbf{x})}[U_\theta(\mathbf{x})] - \mathbb{E}_{p_{\text{data}}(\mathbf{x})}\mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[-\log \underbrace{\int \exp(-U_\theta(\mathbf{x}))q(\mathbf{y}\,|\,\mathbf{x}')\mathrm{d}\mathbf{x}'}_{\mathbb{E}_{q(\mathbf{x}'|\mathbf{y})}[\exp(-U_\theta(\mathbf{x}')]}\right]$$

Data Samples: $\mathbf{x}^i \overset{i.i.d}{\sim} p_{\text{data}}$

$\boldsymbol{\xi}^i \overset{i.i.d.}{\sim} \mathcal{N}(0,\mathbf{I})$

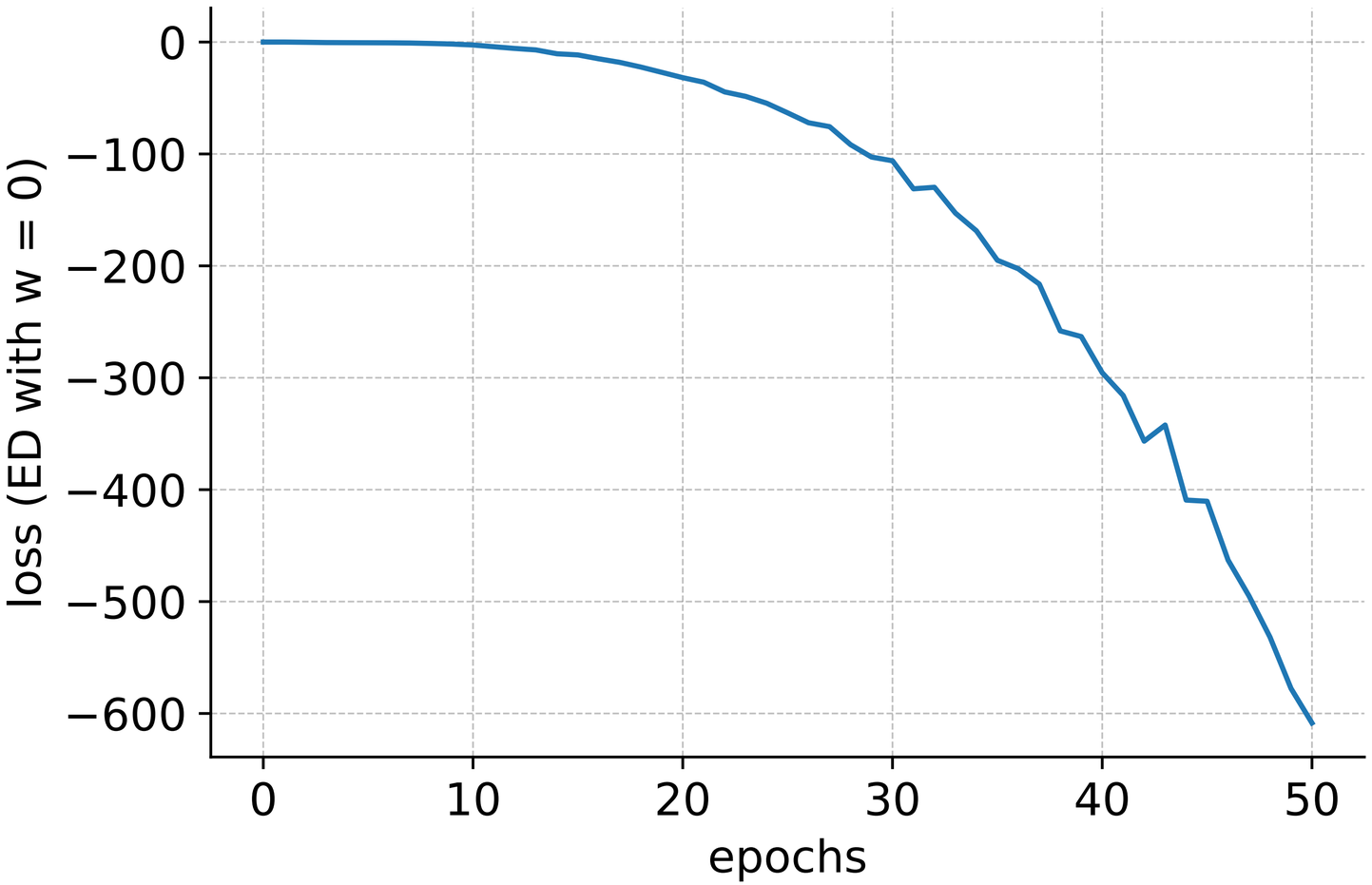
Perturbed Samples: $\mathbf{y}^i = \mathbf{x}^i + \sqrt{t}\boldsymbol{\xi}^i$

$\boldsymbol{\zeta}^{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0,\mathbf{I})$

Contrastive Samples: $\mathbf{x}_-^{i,j} := \mathbf{x}^i + \sqrt{t}\boldsymbol{\xi}^i + \sqrt{t}\boldsymbol{\zeta}^{i,j}$

$$\mathscr{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{1}{M}\sum_{j=1}^{M}\exp(U_\theta(\mathbf{x}^i) - U_\theta(\mathbf{x}_-^{ij}))\right)$$

Looks good but…



What is happening here?

The sample approximation may be unstable

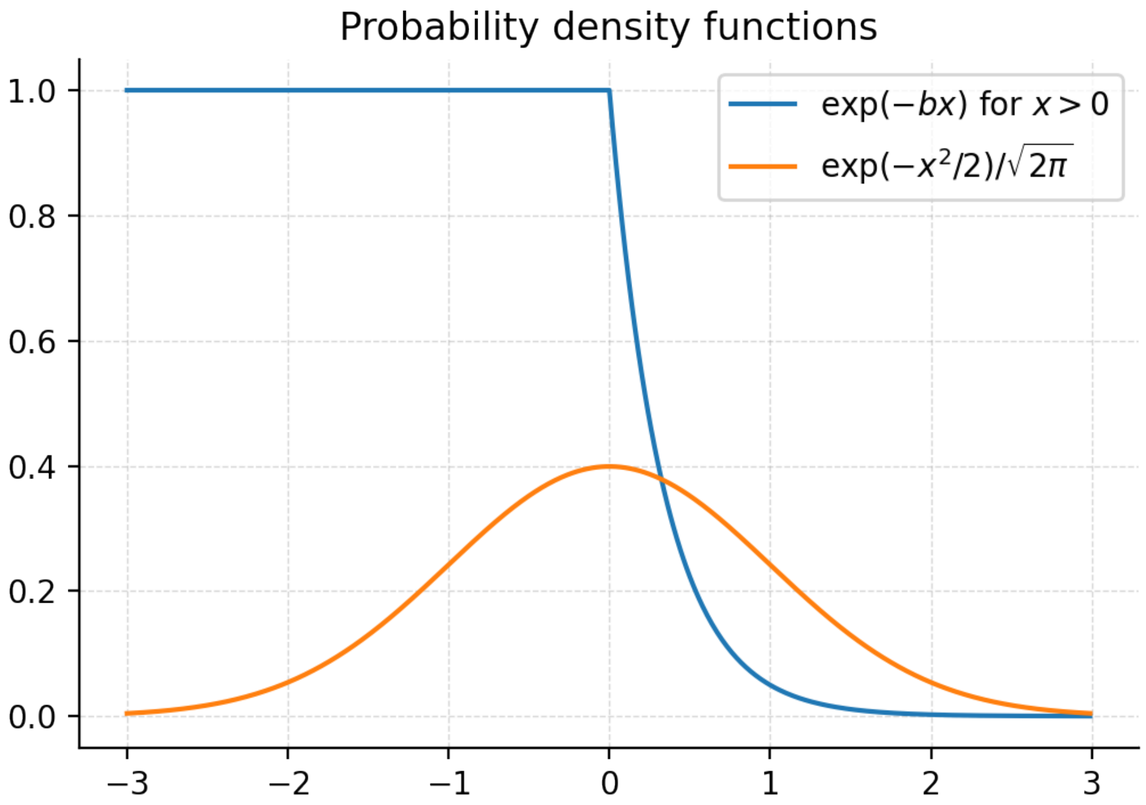
Contrastive potential is bounded

$$U_{\gamma_1}(0) := -\log \int \exp(-U_b(x))\gamma_1(x)\mathrm{d}x < \log(2)$$

Particle approximation is not!

$\xi \sim \mathcal{N}(0,1)$, by chance $\xi > 0$:

$$\widehat{U}_{\gamma_1}(0) = U_b(\xi) = b\xi \xrightarrow{b\to\infty} \infty$$



Stabilising Energy Discrepancy Estimation

$$\mathscr{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{1}{M}\sum_{j=1}^{M} \exp(U_\theta(\mathbf{x}^i) - U_\theta(\mathbf{x}_-^{ij}))\right)$$

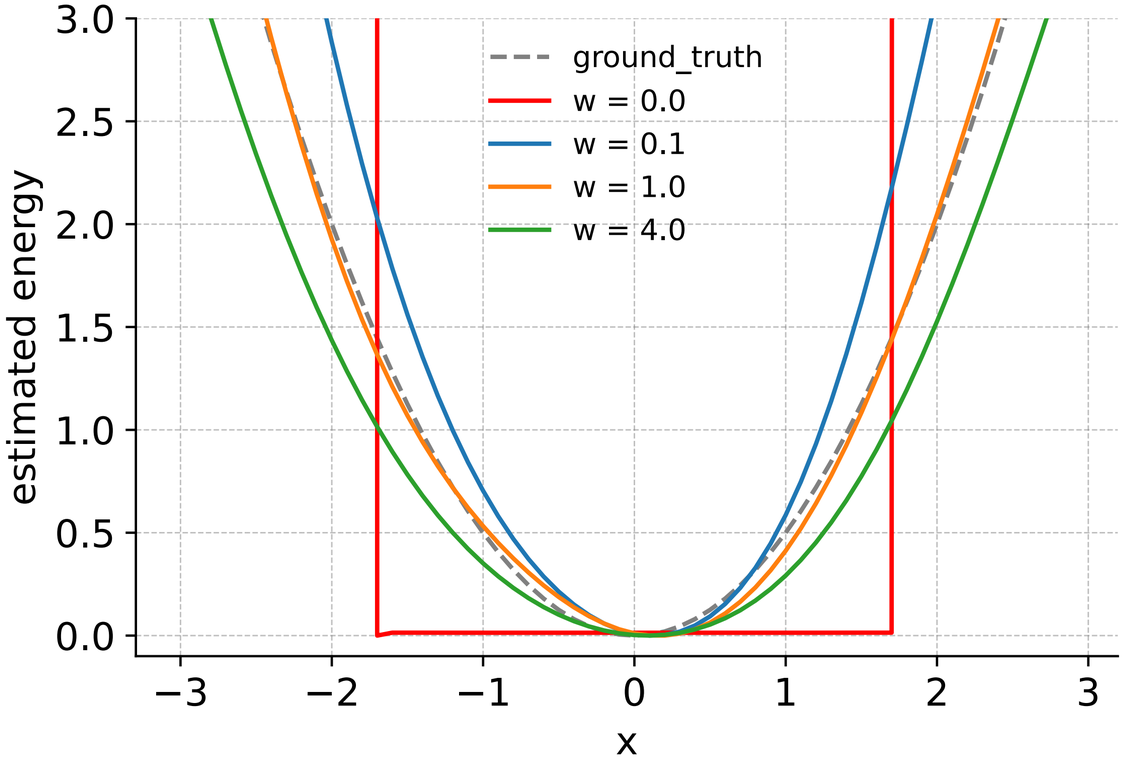
$$\mathscr{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{w}{M} + \frac{1}{M}\sum_{j=1}^{M} \exp(U_\theta(\mathbf{x}^i) - U_\theta(\mathbf{x}_-^{ij}))\right)$$

$$\geq \max\{\log(w), U_\theta(\mathbf{x}^i) - U_\theta(\mathbf{x}_-^{i,1}), \ldots, U_\theta(\mathbf{x}^i) - U_\theta(\mathbf{x}_-^{i,M})\} - \log(M)$$

Terms with $U_\theta(\mathbf{x}_+) - U_\theta(\mathbf{x}_-) < \log(w)$ are dampened

Stabilising Energy Discrepancy Estimation

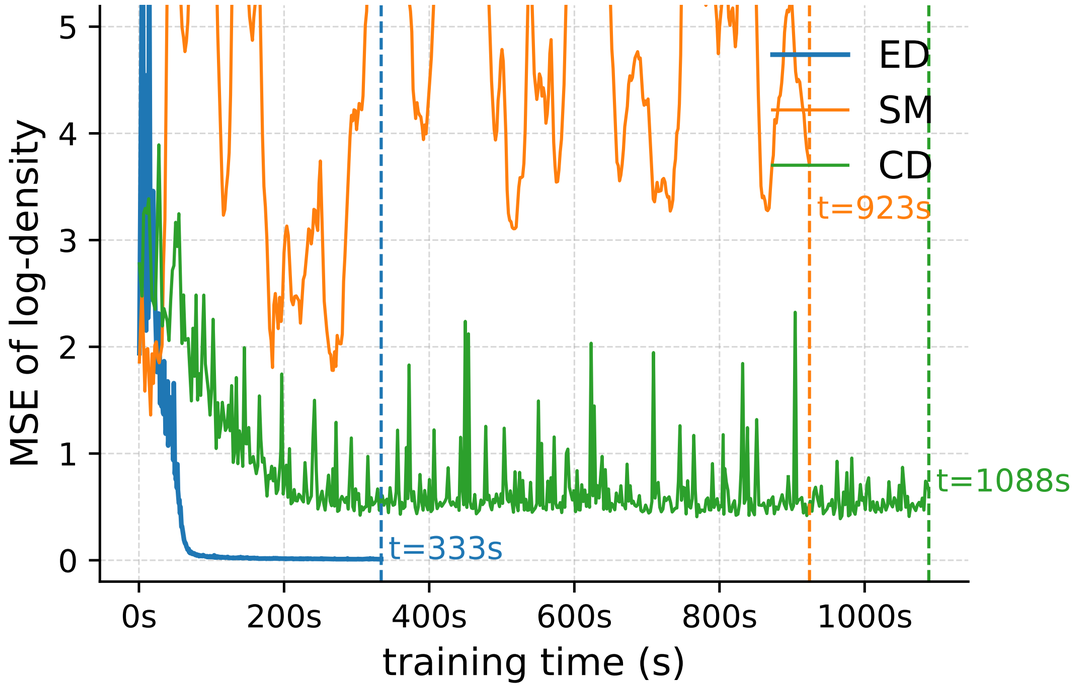
$$\mathscr{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{w}{M} + \frac{1}{M}\sum_{j=1}^{M} \exp(U_\theta(\mathbf{x}^i) - U_\theta(\mathbf{x}_-^{ij}))\right)$$

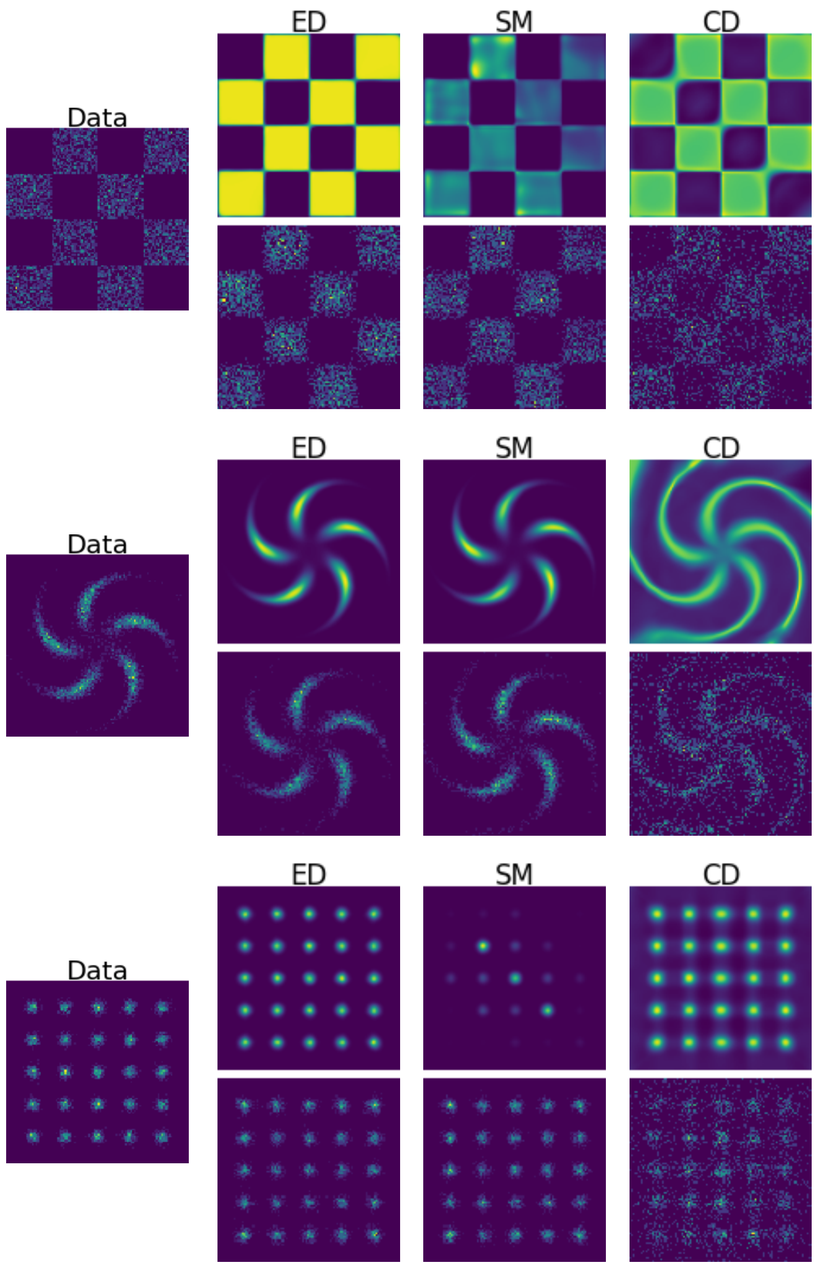


Experimental Results

Density Estimation

$$\mathscr{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{1}{4} + \frac{1}{4}\sum_{j=1}^{4} \exp(U_\theta(\mathbf{x}^i) - U_\theta(\mathbf{x}_-^{ij})\right)$$





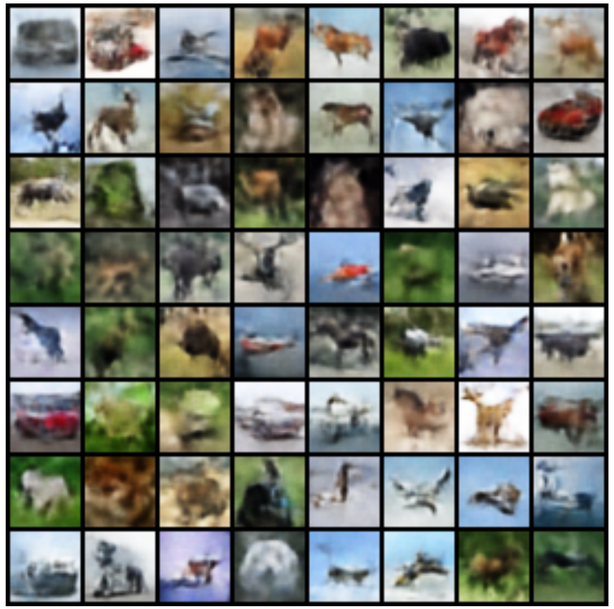
Experimental Results

Image Modelling

- Failure when ED is applied to image data naively due to the manifold hypothesis
- Train latent EBM (Pang et al., 2020) with Energy Discrepancy

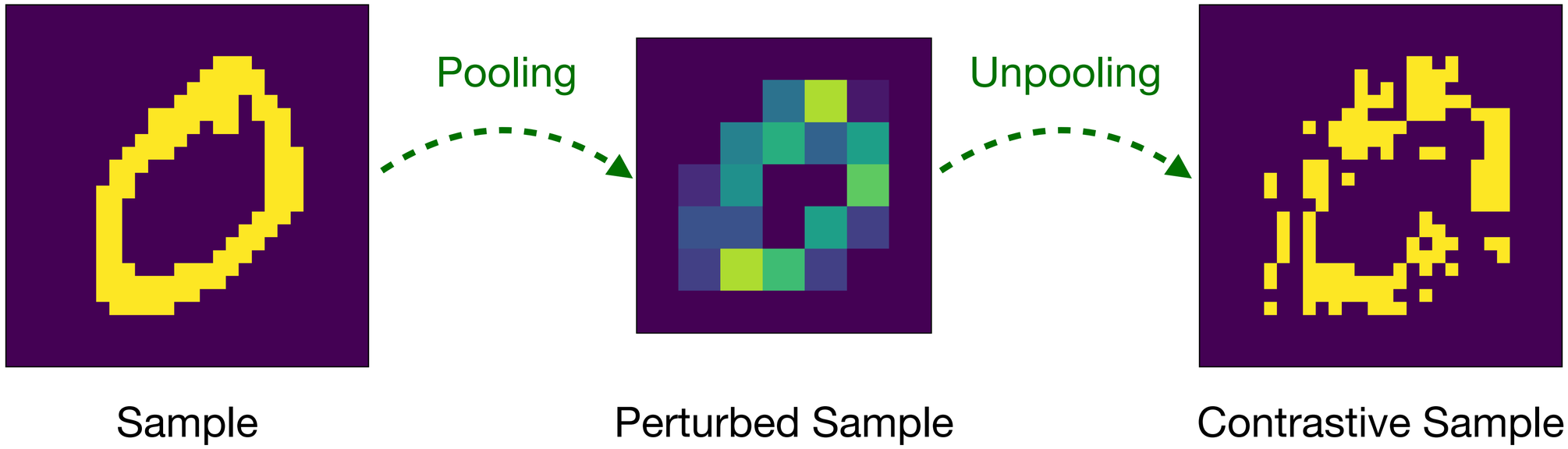
Table 1: Comparison of MSE(↓) and FID(↓) on the SVHN, CIFAR-10, and CelebA datasets.

	SVHN		CIFAR-10		CelebA	
	MSE	FID	MSE	FID	MSE	FID
VAE (Kingma & Welling, 2013)	0.019	46.78	0.057	106.37	0.021	65.75
2s-VAE (Dai & Wipf, 2019)	0.019	42.81	0.056	72.90	0.021	44.40
RAE (Ghosh et al., 2019)	0.014	40.02	0.027	74.16	0.018	40.95
SRI (Nijkamp et al., 2020b)	0.018	44.86	0.020	-	-	61.03
SRI (L=5) (Nijkamp et al., 2020b)	0.011	35.32	-	-	0.015	47.95
CD-LEBM (Pang et al., 2020)	0.008	29.44	**0.020**	**70.15**	0.013	37.87
SM-LEBM	0.010	34.44	0.026	77.82	0.014	41.21
ED-LEBM (ours)	**0.006**	**28.10**	0.023	73.58	**0.009**	**36.73**



Going beyond the Euclidean domain

- Energy-Discrepancy is well-suited for non-Euclidean domains where the computation of scores or MCMC is hard

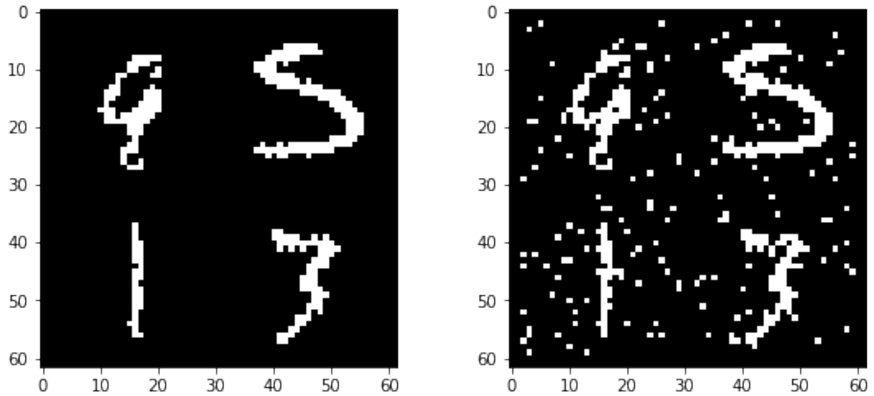


$$\mathscr{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{w}{M} + \frac{1}{M}\sum_{j=1}^{M} \exp(U_\theta(\mathbf{x}_+^i) - U_\theta(\mathbf{x}_-^{ij}))\right)$$

Bernoulli Perturbation

- Consider binary data vectors: $\mathscr{X} = \{0,1\}^d$
- Flip dimensions at entries of Bernoulli distributed mask $\boldsymbol{\xi} \sim \mathrm{Bern}(\epsilon)$

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



Neighbourhood Perturbation

- Define a neighbourhood structure $\mathscr{X} \to \mathscr{X}^K, \mathbf{x} \mapsto \mathscr{N}(\mathbf{x})$
- Perturb by uniformly sampling from neighbourhood:

$$\mathbf{y} \sim \mathrm{Uniform}(\mathscr{N}(\mathbf{x}))$$

- Obtain contrastive samples from inverse neighbourhood:

$$\mathbf{x}\_ \sim \mathscr{N}^{-1}(\mathbf{y})$$

- We typically use the 1-bit neighbourhood

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

-

Deterministic Transformation

- Let g be a deterministic transformation.
- Consider

$$\text{ED}_g(p_{data}, p) = \text{KL}(p_{data}, p) - \text{KL}(g^\sharp p_{data}, g^\sharp p).$$

- If g is invertible, then $\text{KL}(g^\sharp p_{data}, g^\sharp p) = \text{KL}(p_{data}, p)$ —> Not very useful!
- If g loses information, e.g. mean pooling then we can use it!

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

Results in the discrete domain

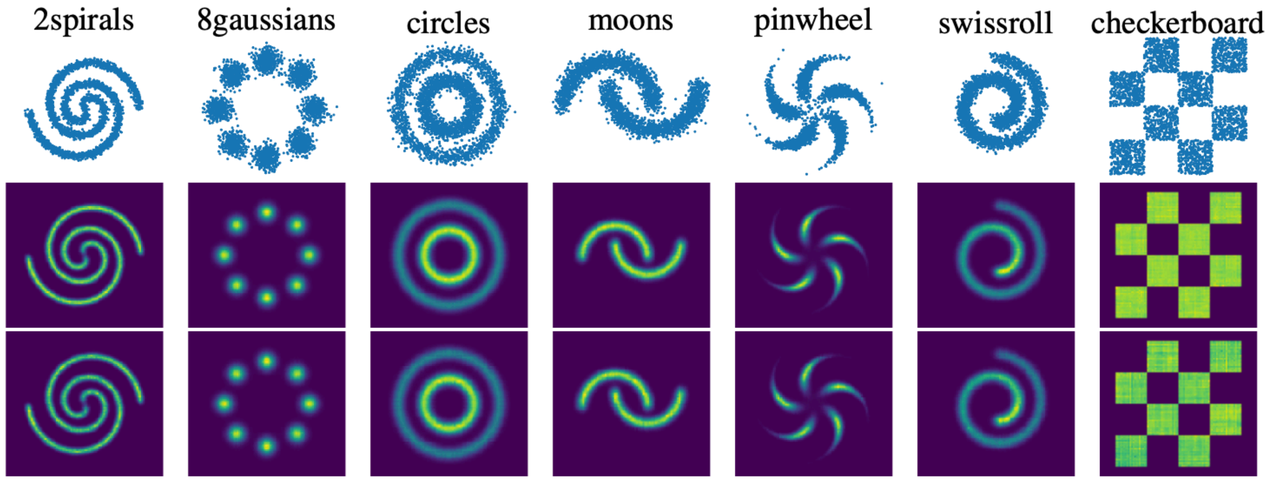
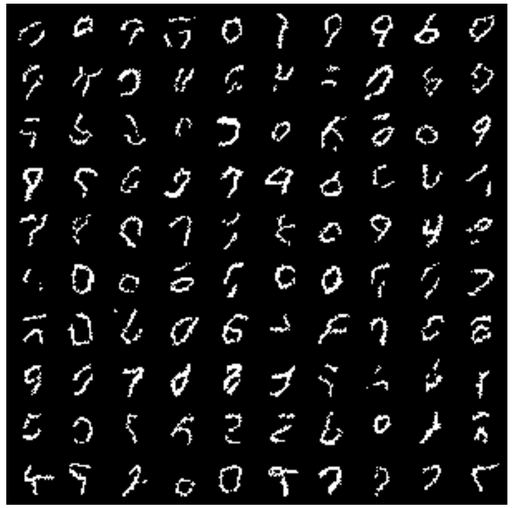
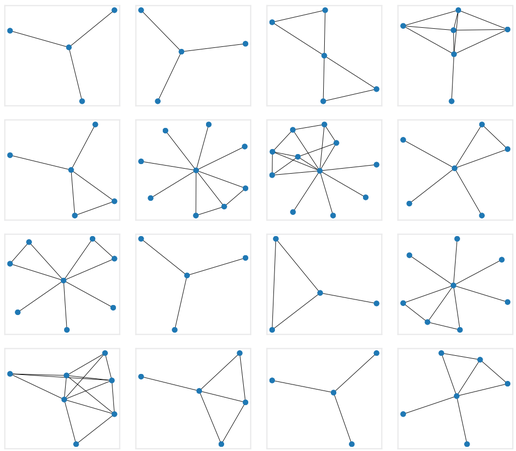


Figure 2: Visualization of training samples and the learned energy landscapes for discrete density estimation. Top to Bottom: training samples, energy landscapes learned by ED-Bern and ED-Grid. We defer more qualitative results to Figure 4.

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-∇Bern	20.048	19.979	20.603	19.717	**19.553**	20.089	**20.677**
	ED-Grid	20.049	**19.965**	20.601	19.715	19.564	20.088	20.678
	ED-∇Grid	20.092	20.005	20.605	19.740	19.577	20.087	21.439





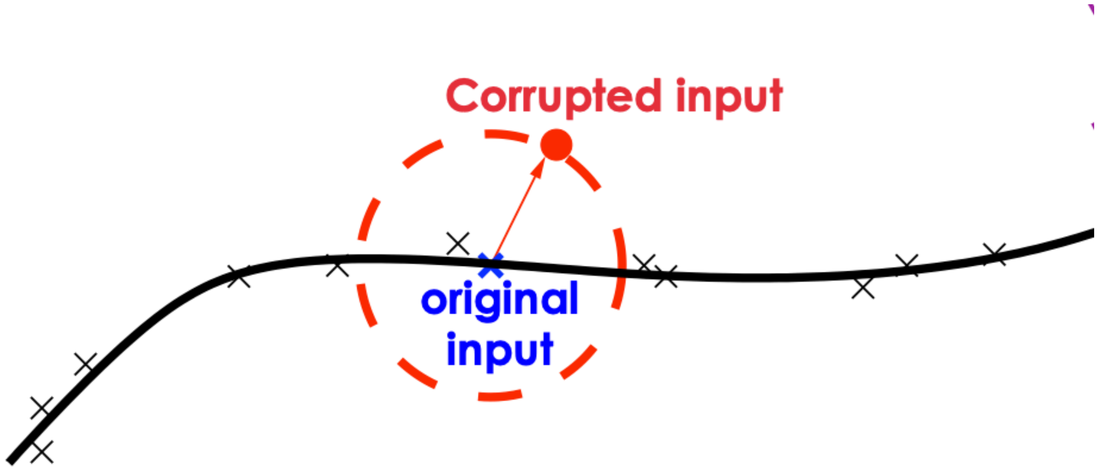
The curse of dimensionality for Energy Discrepancy

- In high dimensions:
  - Data concentrates on a singular support

    ➡ $\log p_{\text{data}}$ undefined almost everywhere

  - Noise is orthogonal to data almost surely

    ➡ Noise uninformative for learning



Bengio et al. (2014)
Representation Learning: A Review and New Perspectives

Thank you!

a.duncan@imperial.ac.uk

github.com/J-zin/energy-discrepancy

References

- Schroeder et al. (2023) Energy Discrepancies: A Score-Independent Loss for Energy-Based Models, In NeurIPS 2023
- Grathwohl et al. (2020) Your Classifier is Secretly an Energy-Based Model and you should treat it like one, In ICLR 2020
- Glaser et al. (2023) Maximum Likelihood Learning of Unnormalized Models for Simulation-Based Inference
- Zhang et al. (2022) Towards Healing the Blindness of Score Matching
- Pang et al. (2020) Learning Latent Space Energy-Based Prior
- Bengio et al. (2014) Representation Learning: A Review and New Perspectives