Implicit Generation and Modeling with Energy-Based Models

Yilin Du, Igor Mordatch, 2019

Brenda Huppenthal, 3/27/2024

Contributions

- Present an algorithm (and in the supplemental, techniques) for training energy based models (EBMs) on high dimensional data
- Present empirical results on compositionality, decorruption, inpainting
- Show that EBMs are useful in a wide variety of domains like out of distribution detection, adversarially robust classification, trajectory prediction, online learning

Energy Based Models

Use a deep neural network (parameterized by θ) to learn an **energy function**:

$$E_\theta(x) \in \mathbb{R}$$

This energy function defines a probability distribution function via the **Boltzmann distribution**:

$$p_\theta(x) = \frac{exp\left[-E_\theta(x)\right]}{\int exp\left[-E_\theta(x)\right]dx} = \frac{exp\left[-E_\theta(x)\right]}{Z(\theta)}$$

Here, Z(θ) is the **partition function** and is intractable.

Sampling

Generating samples from this distribution is challenging. Previous methods used MCMC methods like random walk and Gibbs sampling, which both suffer from long mixing times especially for high-dimensional data like images.

We can speed this up using **Langevin dynamics**, which uses the gradient of the energy function.

$$\tilde{x}^k = \tilde{x}^{k-1} - \frac{\lambda}{2}\nabla_x E_\theta\left(\tilde{x}^{k-1}\right) + \omega^k \qquad \omega^k \sim \mathcal{N}(0, \lambda)$$

This sampling procedure defines a distribution q_θ: $x^k \sim q_\theta$.

Sampling

drift diffusion

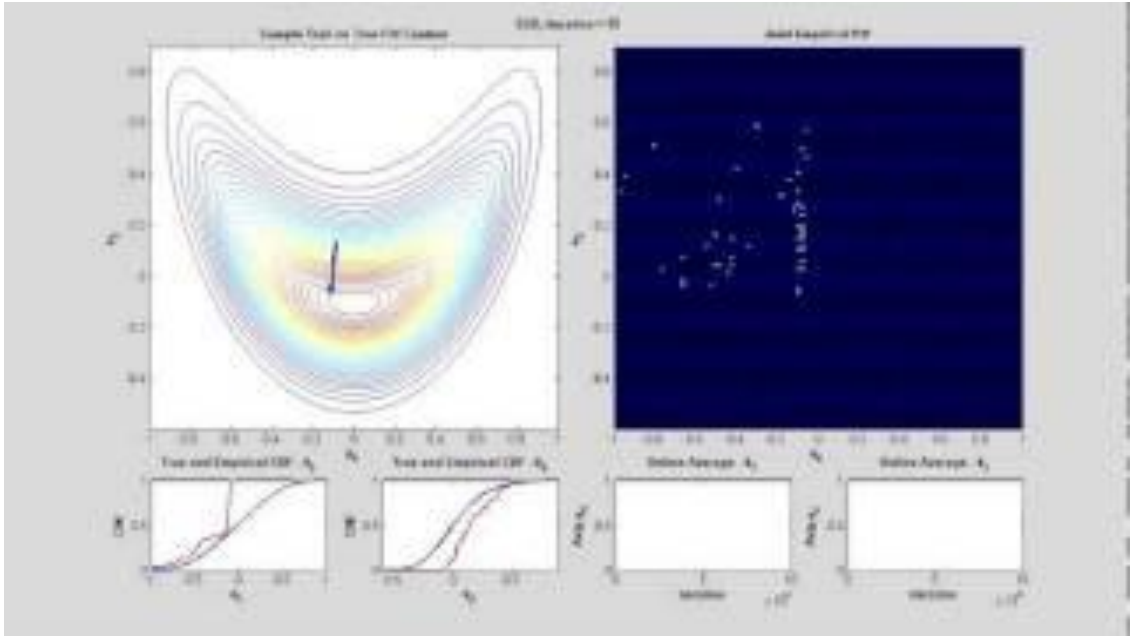
$$\tilde{x}^k = \tilde{x}^{k-1} - \frac{\lambda}{2}\nabla_x E_\theta\left(\tilde{x}^{k-1}\right) + \omega^k \qquad \omega^k \sim \mathcal{N}(0, \lambda)$$

Welling and Teh use Langevin dynamics to sample from the true posterior distribution while performing stochastic gradient descent.

They proved that as $K \to \infty$ and $\boldsymbol{\lambda} \to 0$, $q_\theta \to p_\theta$.

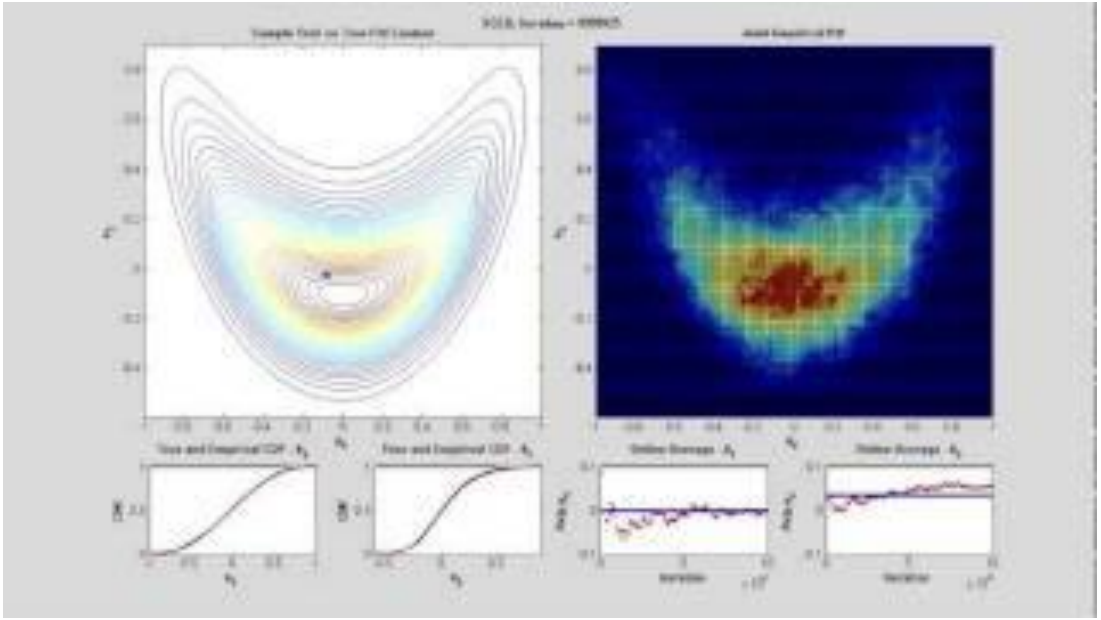
They note that without the added noise, this would collapse to the nearest MAP solution.

Stochastic Gradient Descent



Welling and Teh: Bayesian Learning via Stochastic Gradient Langevin Dynamics

Stochastic Gradient Langevin Sampling



Welling and Teh: Bayesian Learning via Stochastic Gradient Langevin Dynamics

Energy Based Models: Two Views

1. Defines a probability distribution over the data

$$p_\theta(x) = \frac{exp\left[-E_\theta(x)\right]}{Z(\theta)}$$

2. Defines an implicit generator

$$\tilde{x}^k = \tilde{x}^{k-1} - \frac{\lambda}{2}\nabla_x E_\theta\left(\tilde{x}^{k-1}\right) + \omega^k \qquad \omega^k \sim \mathcal{N}(0, \lambda)$$

Maximum Likelihood Training

We want to push the distribution defined by our energy function E to model the data distribution:

$$\mathcal{L}_{ML}(\theta) = \mathbb{E}_{x \sim p_D}\left[-\log p_\theta(x)\right] = \mathbb{E}_{x \sim p_D}\left[E_\theta(x) - \log Z(\theta)\right]$$

Turner 2005 derives the gradient of this loss function:

$$\nabla_\theta \mathcal{L}_{ML} = \mathbb{E}_{x^+ \sim p_D}\left[\nabla_\theta E_\theta(x^+)\right] - \mathbb{E}_{x^- \sim p_\theta}\left[\nabla_\theta E_\theta(x^-)\right]$$

$$\approx \mathbb{E}_{x^+ \sim p_D}\left[\nabla_\theta E_\theta(x^+)\right] - \mathbb{E}_{x^- \sim q_\theta}\left[\nabla_\theta E_\theta(x^-)\right]$$

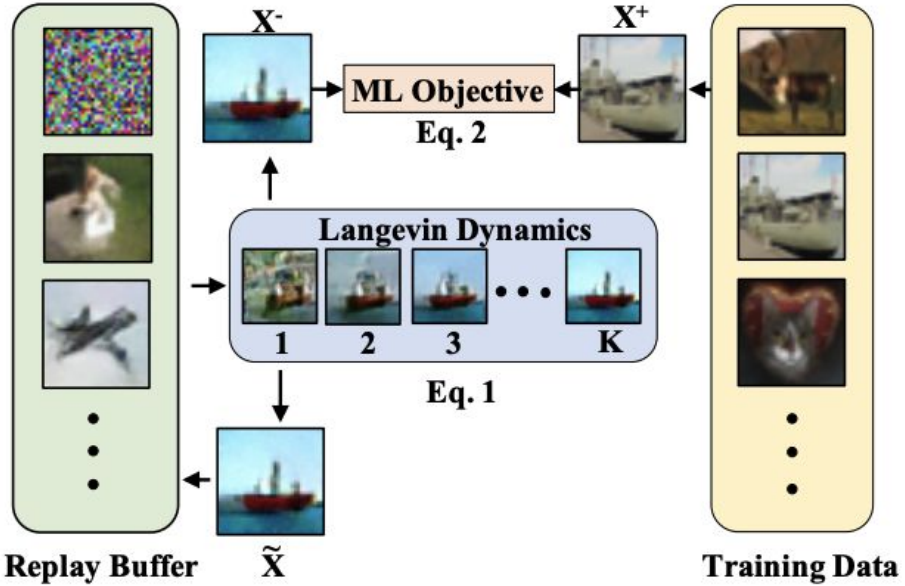
Minimize the energy of positive samples from the data

Maximize the energy of negative (hallucinated) generated samples

Sample Replay Buffer

Sample replay buffer $\mathcal{B}$ holds previously generated samples. These can be used to initialize the Langevin dynamics procedure. With 95% probability, pick a sample from the replay buffer, else use uniform noise.

Because the sampling procedure is a Markov chain, this gives us a headstart on mixing time and reduces training time.

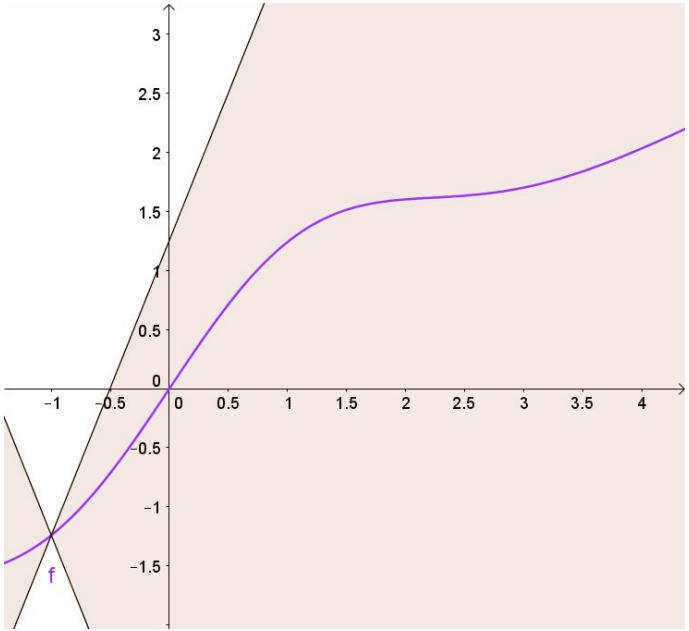


Regularization

Arbitrary energy functions can have very sharp changes in the gradient that make Langevin dynamics unstable, and thus makes training and generation difficult.

Constraining the Lipschitz constant of the energy function helps these issues, which they do by adding spectral normalization to all layers of the energy model.

They also add weak L2 regularization of energy magnitudes.



Algorithm

Algorithm 1 Energy training algorithm

Input: data dist. $p_D(\mathbf{x})$, step size λ, number of steps K

$\mathcal{B} \leftarrow \varnothing$

while not converged **do**

$\mathbf{x}_i^+ \sim p_D$

$\mathbf{x}_i^0 \sim \mathcal{B}$ with 95% probability and $\mathcal{U}$ otherwise

▷ *Generate sample from q_θ via Langevin dynamics:*

for sample step $k = 1$ to K **do**

$\tilde{\mathbf{x}}^k \leftarrow \tilde{\mathbf{x}}^{k-1} - \nabla_{\mathbf{x}} E_\theta(\tilde{\mathbf{x}}^{k-1}) + \omega, \quad \omega \sim \mathcal{N}(0, \sigma)$

end for

$\mathbf{x}_i^- = \Omega(\tilde{\mathbf{x}}_i^k)$

▷ *Optimize objective $\alpha\mathcal{L}_2 + \mathcal{L}_{ML}$ wrt θ:*

$\Delta\theta \leftarrow \nabla_\theta \frac{1}{N} \sum_i \alpha(E_\theta(\mathbf{x}_i^+)^2 + E_\theta(\mathbf{x}_i^-)^2) + E_\theta(\mathbf{x}_i^+) - E_\theta(\mathbf{x}_i^-)$

Update θ based on $\Delta\theta$ using Adam optimizer

$\mathcal{B} \leftarrow \mathcal{B} \cup \tilde{\mathbf{x}}_i$

end while

Image Generation

(a) GLOW Model (b) EBM (c) EBM (10 historical) (d) EBM Sample Buffer

Figure 3: Comparison of image generation techniques on unconditional CIFAR-10 dataset.

Image Generation

Model	Inception*	FID
CIFAR-10 Unconditional		
PixelCNN [Van Oord et al., 2016]	4.60	65.93
PixelIQN [Ostrovski et al., 2018]	5.29	49.46
EBM (single)	6.02	40.58
DCGAN [Radford et al., 2016]	6.40	37.11
WGAN + GP [Gulrajani et al., 2017]	6.50	36.4
EBM (10 historical ensemble)	6.78	38.2
SNGAN [Miyato et al., 2018]	**8.22**	21.7
CIFAR-10 Conditional		
Improved GAN	8.09	-
EBM (single)	8.30	37.9
Spectral Normalization GAN	**8.59**	25.5
ImageNet 32x32 Conditional		
PixelCNN	8.33	33.27
PixelIQN	10.18	22.99
EBM (single)	**18.22**	**14.31**
ImageNet 128x128 Conditional		
ACGAN [Odena et al., 2017]	28.5	-
EBM* (single)	28.6	43.7
SNGAN	**36.8**	**27.62**

Figure 4: Table of Inception and FID scores for ImageNet32x32 and CIFAR-10. Quantitative numbers for ImageNet32x32 from [Ostrovski et al., 2018]. (*) We use Inception Score (from original OpenAI repo) to compare with legacy models, but strongly encourage future work to compare soley with FID score, since Langevin Dynamics converges to minima that artificially inflate Inception Score. (**) conditional EBM models for 128x128 are smaller than those in SNGAN.

Denoising and Inpainting

Figure 5: EBM image restoration on images in the **test** set via MCMC. The right column shows failure (approx. 10% objects change with ground truth initialization and 30% of objects change in salt/pepper corruption or inpainting. Bottom two rows shows worst case of change.)

Mode Coverage

Figure 6: Illustration of cross-class implicit sampling on a conditional EBM. The EBM is conditioned on a particular class but is initialized with an image from a separate class.

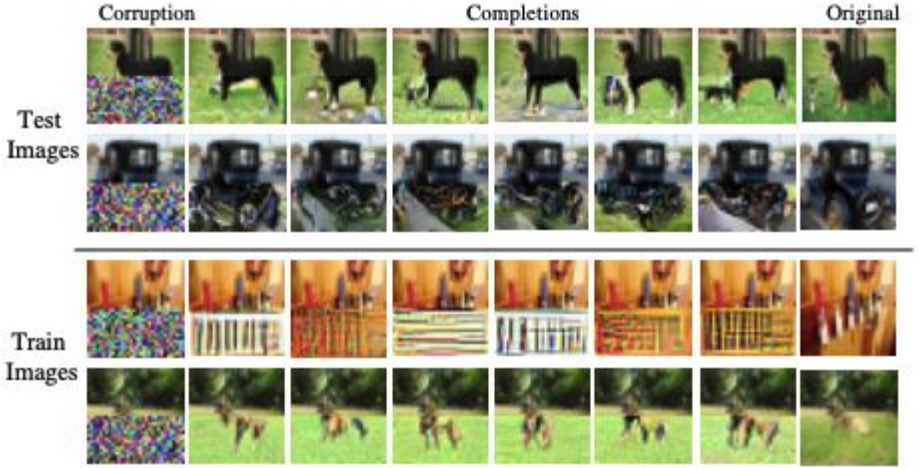


Figure 7: Illustration of image completions on conditional ImageNet model. Our models exhibit diversity in inpainting.

Out-of-Distribution Detection

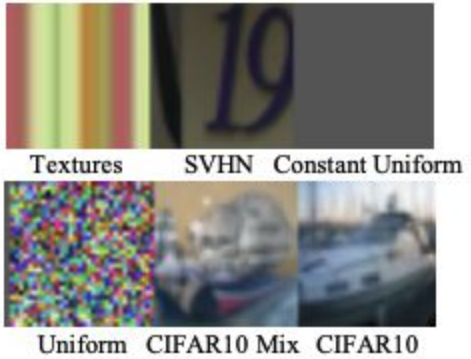


Figure 9: Illustration of images from each of the out of distribution dataset.

Model	PixelCNN++	Glow	EBM (ours)
SVHN	0.32	0.24	**0.63**
Textures	0.33	0.27	**0.48**
Constant Uniform	0.0	0.0	**0.30**
Uniform	1.0	1.0	**1.0**
CIFAR10 Interpolation	**0.71**	0.59	0.70
Average	0.47	0.42	**0.62**

Figure 10: AUROC scores of out of distribution classification on different datasets. Only our model gets better than chance classification.

Out-of-Distribution Generalization

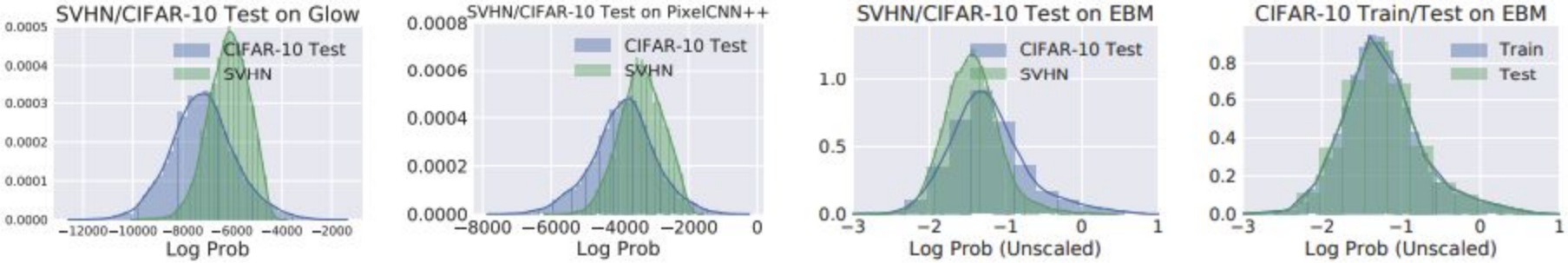
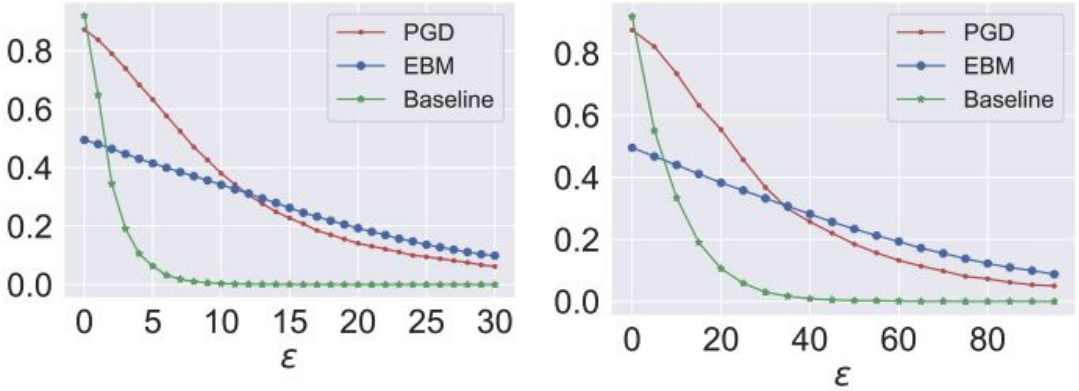


Figure 11: Histogram of relative likelihoods for various datasets for Glow, PixelCNN++ and EBM models

Adversarial Robustness



(a) L_∞ robustness (b) L_2 Robustness

Figure 8: ϵ plots under L_∞ and L_2 attacks of conditional EBMs as compared to PGD trained models in [Madry et al., 2017] and a baseline Wide ResNet18.

Trajectory Modeling

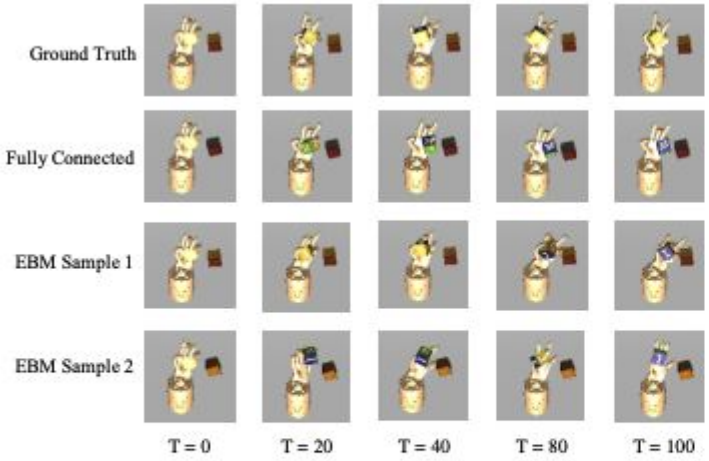


Figure 12: Views of hand manipulation trajectories generated unconditionally from the same state(1st frame).

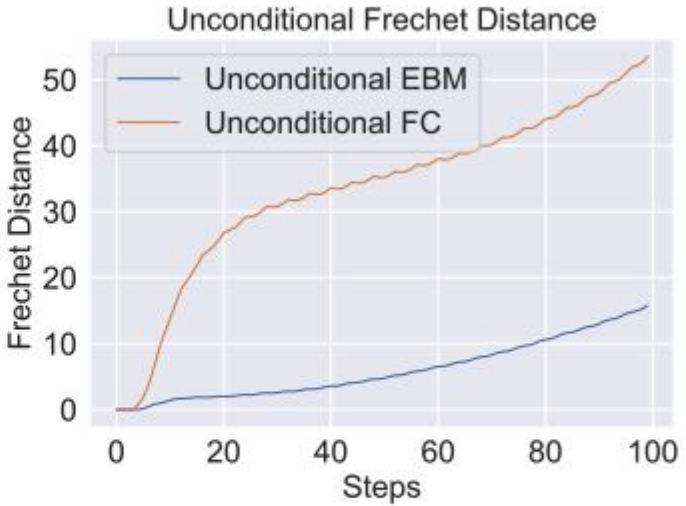


Figure 13: Conditional and Unconditional Modeling of Hand Manipulation through Frechet Distance

Online Learning

Method	Accuracy
EWC [Kirkpatrick et al., 2017]	19.80 (0.05)
SI [Zenke et al., 2017]	19.67 (0.09)
NAS [Schwarz et al., 2018]	19.52 (0.29)
LwF [Li and Snavely, 2018]	24.17 (0.33)
VAE	40.04 (1.31)
EBM (ours)	**64.99** (4.27)

Table 1: Comparison of various continual learning benchmarks. Values averaged acrossed 10 seeds reported as mean (standard deviation).

Compositional Generation

We can compose different EBMs through summation.

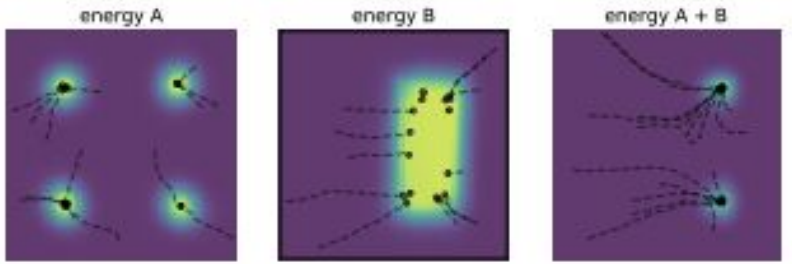


Figure 14: A 2D example of combining EBMs through summation and the resulting sampling trajectories.

Compositional Generation

Sampling a joint distribution on multiple latents is equivalent to generation on a sum of conditional EBMs.

Figure 15: Samples from joint distribution of 4 independent conditional EBMs on scale, position, rotation and shape (left panel) with associated ground truth rendering (right panel).

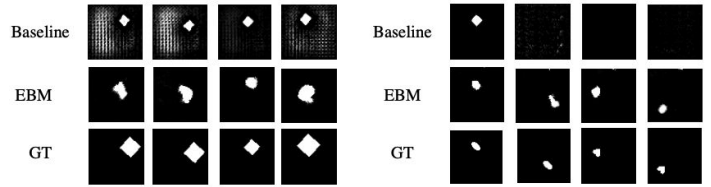


Figure 16: GT = Ground Truth. Images of cross product generalization of size-position (left panel) and shape-position (right panel).