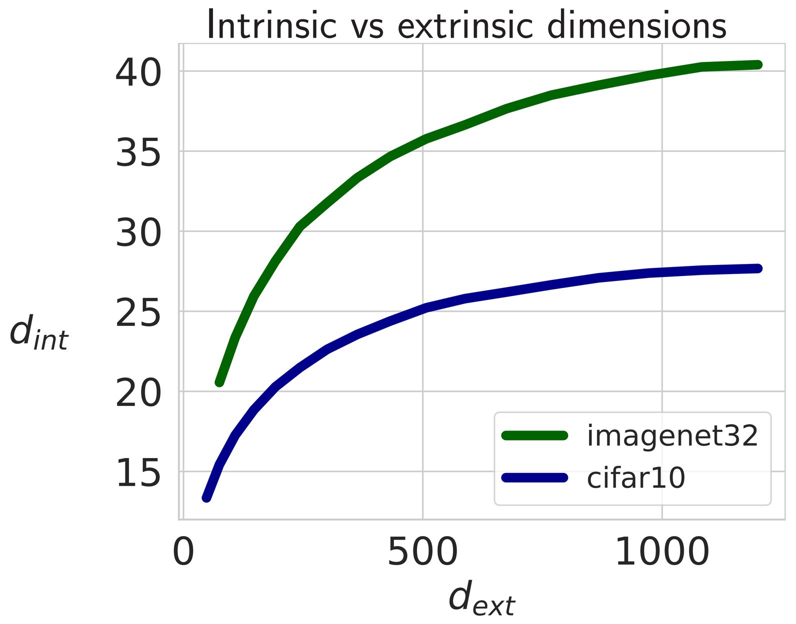
Generalized Energy Based Models

Michael Arbel[1], Liang Zhou[1], and Arthur Gretton[1]
[1]Gatsby Computational Neuroscience Unit, University College London

Overview

Problem

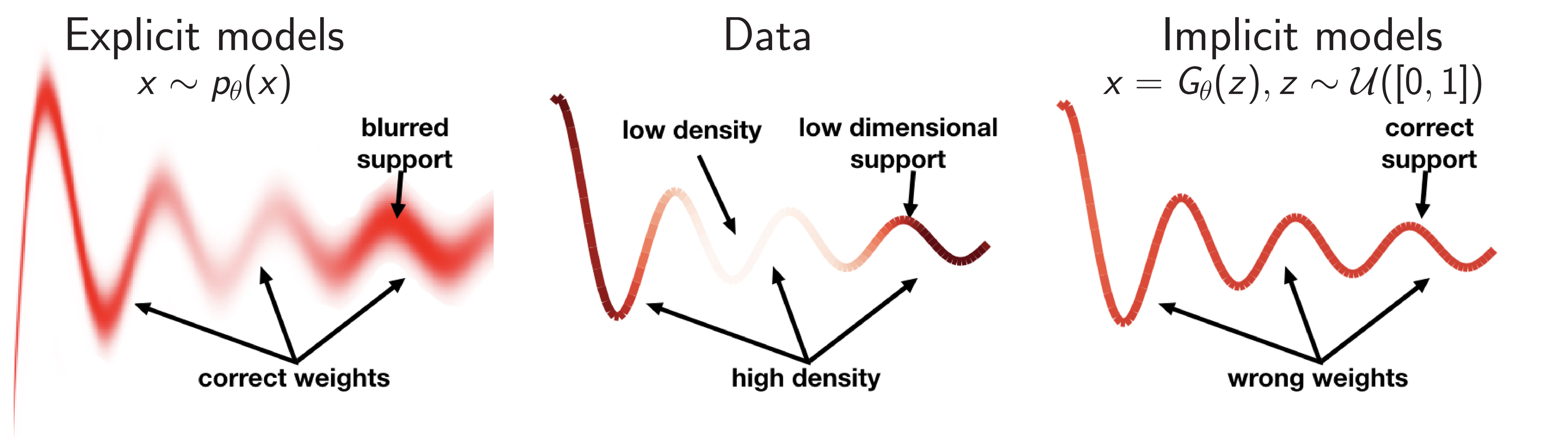
✓ Setting: data distributions with *small intrinsic dimension* embedded in a space with *high extrinsic dimension*.
✓ Example: Includes data such as natural images [2].
✓ Goal: Flexible models exploiting low intrinsic dimensionality.



Contributions

✓ A model with implicit and explicit component for data with low intrinsic dimension.
✓ End-to-end training procedure based on adversarial training.
✓ Sampling using latent space MCMC.

Motivation: Explicit vs Implicit models



- An expressive *implicit model* can recover the weights, but training is waistful: it throws away the critic.
- An *explicit model* puts mass on the whole space: it blurs the samples.

Generalized Energy Based Models (GEBM)s

GEBMs are defined by a combination of the two components: *energy* and *base*

- The base $\mathbb{G}_\theta$ is defined by a fixed latent noise $Z \sim \eta$ pushed-forward by a generator $G_\theta(Z)$.

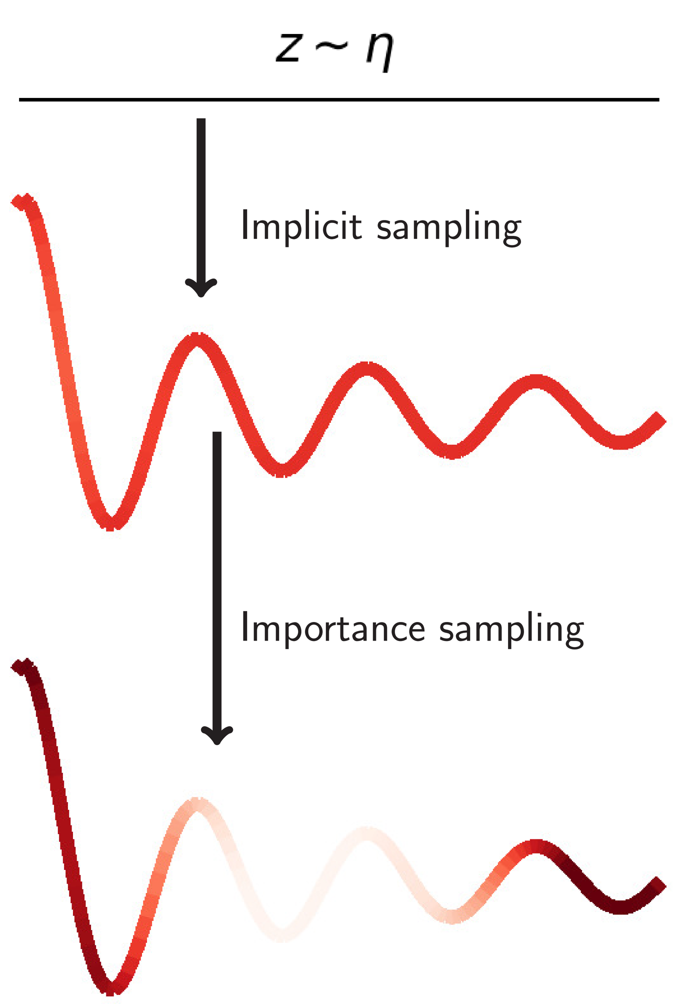
$$X \sim \mathbb{G}_\theta, \quad \Longleftrightarrow \quad X = G_\theta(Z), \quad Z \sim \eta$$

- The base learns the low-dimensional support of the data.
- The energy E defines importance weights on the support of $\mathbb{G}_\theta$

$$w(X) = Z_{\theta,E}^{-1} \exp(-E(X))$$
$$Z_{\theta,E} = \mathbb{E}_{X \sim \mathbb{G}_\theta}[\exp(-E(X))].$$

- The energy *refines* the mass on the low-dimensional support of the base.
- The GEBM $\mathbb{Q}_{\theta,E}$

$$\mathrm{d}\mathbb{Q}_{\theta,E}(X) = w(X)\,\mathrm{d}\mathbb{G}_\theta(X).$$



Learning GEBMs adversarially

Learning the energy

✓ MLE on the support of the base $\mathbb{G}_\theta$

$$\mathcal{L}_{\mathbb{P},\mathbb{G}_\theta}(E) = -\mathbb{E}_{\mathbb{P}}[E] - \log Z_{\theta,E}$$

✓ Amortized estimation of $\log Z_{\theta,E}$ using convexity lower-bound

$$\mathcal{L}_{\mathbb{P},\mathbb{G}_\theta}(E) \geq \underbrace{-\mathbb{E}_{\mathbb{P}}[E+c] - \mathbb{E}_{\mathbb{G}_\theta}\left[e^{-(E+c)}\right]}_{\mathcal{F}_{\mathbb{P},\mathbb{G}_\theta}(E,c)}$$

✓ Tight bound whenever $c = \log Z_{\theta,E}$.

Learning the base

✓ Minimize the KL Approximate Lower-bound Estimator

$$KALE(\mathbb{P}|\mathbb{G}_\theta) = \max_{E,c} \mathcal{F}_{\mathbb{P},\mathbb{G}_\theta}(E,c) := \mathcal{K}(\theta)$$

✓ Well-defined even when data and model have disjoint support
✓ Provably well-defined gradient when energies are L-Lipschitz

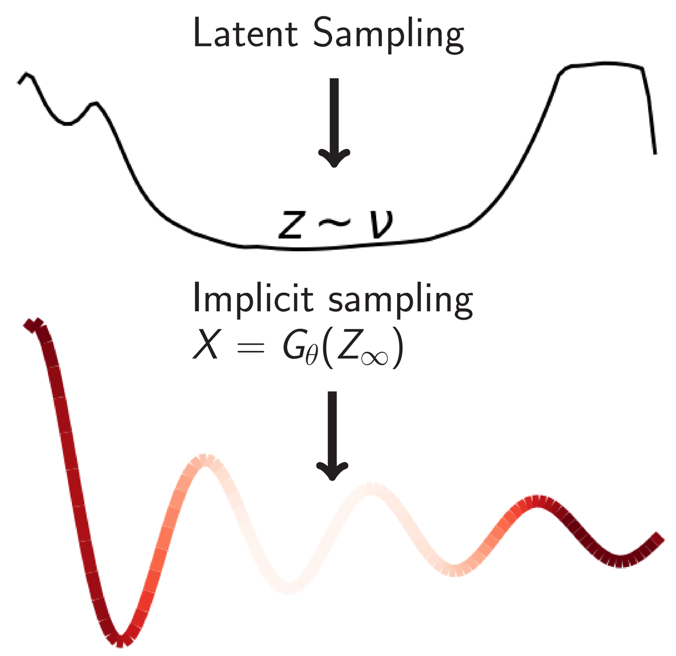
Theorem

If the set of energies is parametrized by a compact set, the energies and their gradients w.r.t. x are L-Lipschitz in x and the generator and its gradient w.r.t. θ are locally L-Lipschitz in θ, then sub-gradient methods on $\mathcal{K}$ converge to local optima. Moreover, $\mathcal{K}$ is Lipschitz and differentiable for almost all $\theta \in \Theta$ with:

$$\nabla\mathcal{K}(\theta) = Z_{\theta,E^\star}^{-1} \int \nabla_x E^\star(G_\theta(z)) \nabla_\theta G_\theta(z) \exp(-E^\star(G_\theta(z)))\eta(z)\,\mathrm{d}z.$$

Sampling from GEBMs using latent MCMC

GEBMs also defined by a learned latent noise $Z \sim \eta(Z)w(G_\theta(Z))$ mapped by G_θ.



- Latents are sampled according to a 'posterior' distribution:

$$\nu(Z) = \eta(Z)w(G_\theta(Z))$$

- Can use Langevin in latent space:

$$W_{k+1} \sim \mathcal{N}(0, I)$$
$$Z_{k+1} = Z_k + \gamma\nabla_z \log\nu(Z_k) + \sqrt{2\gamma}W_{k+1}$$

- Latents are mapped to sample space using the implicit map G_θ:

$$X = G_\theta(Z)$$

Theorem

Assume that $\log\eta(z)$ is strongly concave and has a Lipschitz gradient, that E, $\mathbb{G}_\theta$ and their gradients are all L-Lipschitz. Set $x_t = \mathbb{G}_\theta(z_t)$, where z_t is given by

$$\mathrm{d}z_t = \nabla\log\nu(z_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}w_t,$$

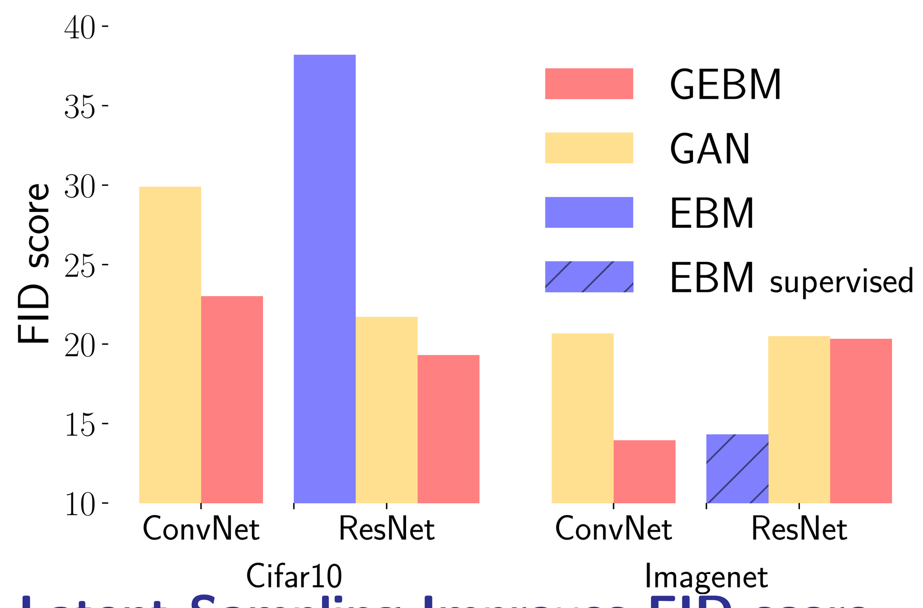
with w_t a Brownian motion. Then $\mathbb{P}_t$, the probability distribution of x_t, converges to $\mathbb{Q}_{\theta,E}$ in the Wasserstein sense,

$$W_2(\mathbb{P}_t, \mathbb{Q}_{\theta,E}) \leq LCe^{-ct},$$

where $c = O(\exp(-dim(\mathcal{Z})))$.

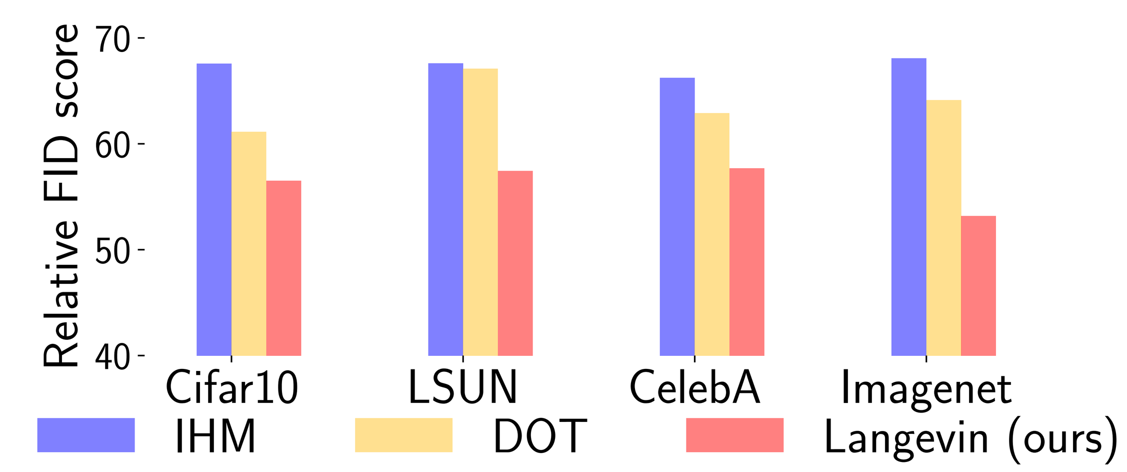
Numerical results

GEBM vs GANs and EBMs



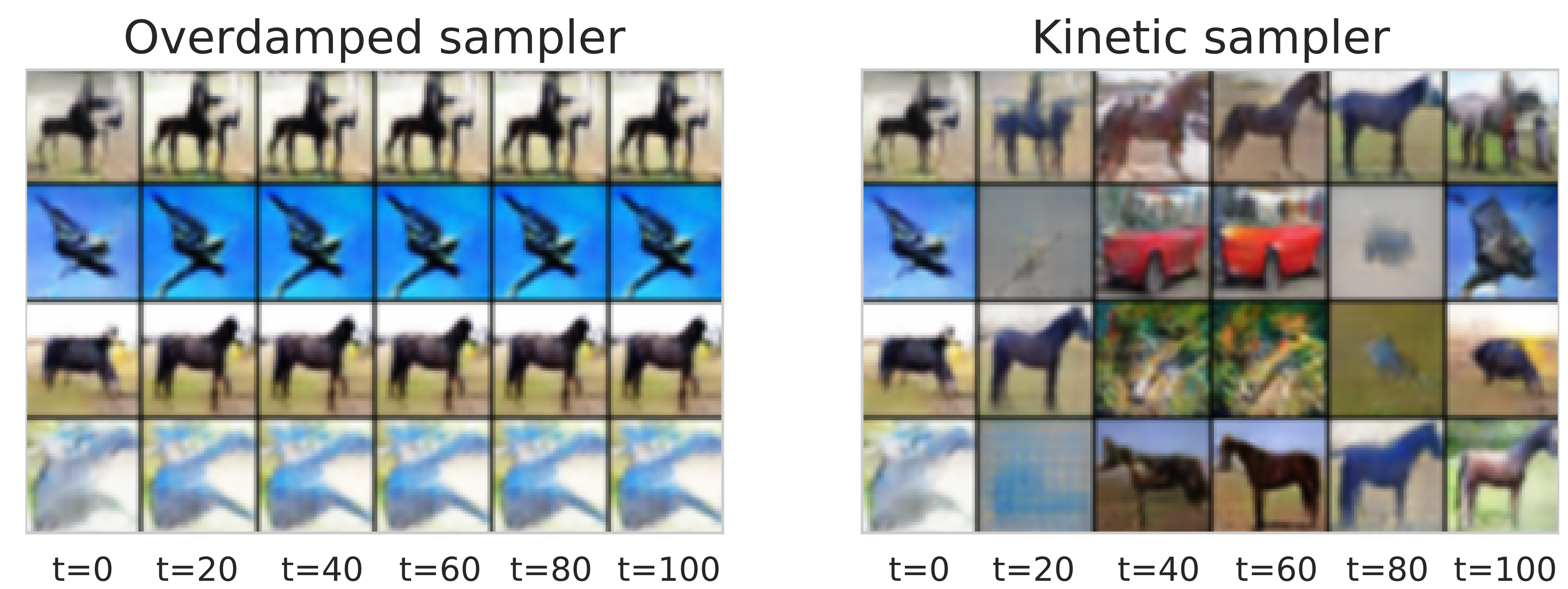
- Using the same models GEBMs outperform GANs:
- The critic/energy contains useful information for sampling.
- GEBMs outperform EBM:
- GEBM exploits the low dimensional assumption as an inductive bias.

Latent Sampling Improves FID score



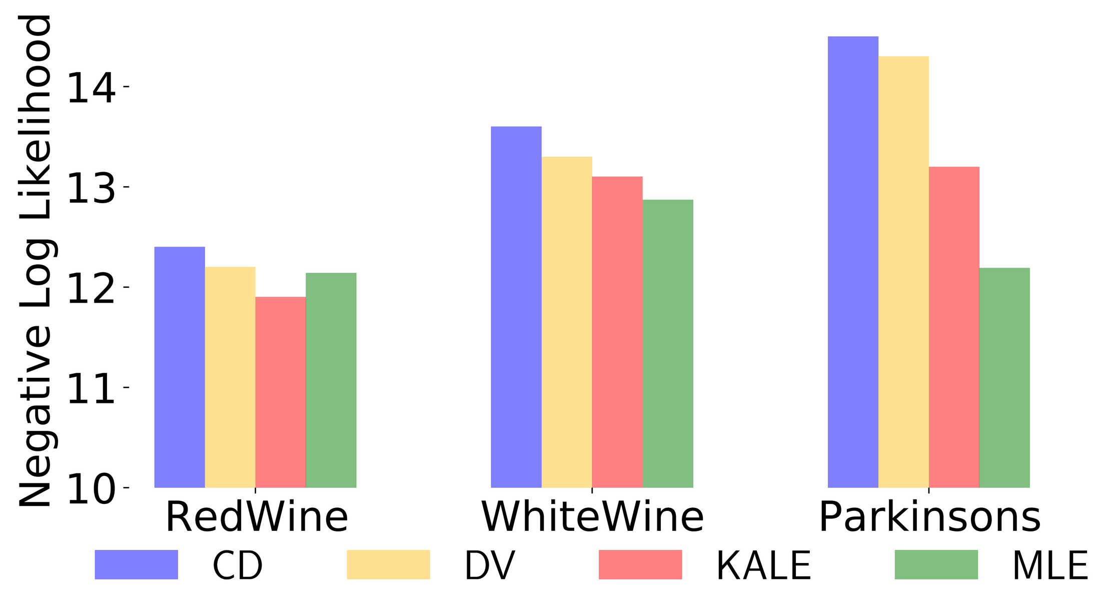
Langevin latent sampling improves over MCMC methods that are not using gradient information the energy (IHM) [3]) and Discriminator Optimal Transport [1].

Overdamped vs kinetic samplers for latent sampling



- Overdamped samplers (like ULA) stick to one particular mode with each chain.
- Kinetic samplers (like HMC) tend to explore multiple modes within the same chain.

Density estimation using adversarial training



- Can use GEBM trained with KALE for density estimation.
- When the base is an NVP, GEBM is an EBM with a learnable ref. measure $\mathbb{G}_\theta$.
- Performance similar to MLE.

Bibliography

A. Tanaka. "Discriminator optimal transport". *Advances in Neural Information Processing Systems*. 2019.
L. Thiry, M. Arbel, E. Belilovsky, and E. Oyallon. "The Unreasonable Effectiveness of Patches in Deep Convolutional Kernels Methods". *ICLR*. 2021.
R. Turner, J. Hung, E. Frank, Y. Saatchi, and J. Yosinski. "Metropolis-hastings generative adversarial networks". *ICML*. 2019.

This work was funded by the Gatsby Charitable Foundation.

Contact: michael.n.arbel@gmail.com