Multimeasurement Generative Models

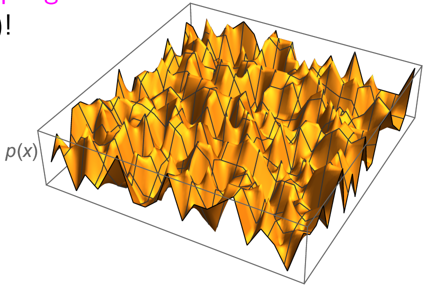
Saeed Saremi[1,2] Rupesh Kumar Srivastava[1]

[1]NNAISENSE, Inc. [2]UC Berkeley

ICLR, 2022

The Big Picture

- The setup: $\{x_i\}_{i=1}^n$ drawn from an *unknown* distribution with density p_X in $\mathbb{R}^d$.
- **Generative modeling is the task of drawing independent samples from p_X.**
- The distributions of interest are typically very "complex".
- The ambient dimension d is typically large $d \gg 1$ ($\approx 2 \times 10^5$ for FFHQ-256).
- Curse of dimensionality: kernel density estimation is doomed in high dimensions.
- Curse of MCMC sampling
  (even if we *knew* p_X)!



- **Our solution: bypass learning/sampling p_X;**
  study a (much) smoother density in $\mathbb{R}^{Md}$.

M-Density

(a) x

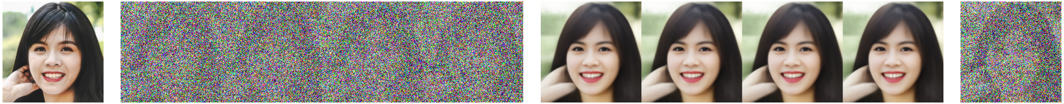
(b) $\mathbf{y} := (y_1, \ldots, y_M)$

- We smooth the (unknown) density p_X with a *factorial* kernel with M channels. The new smoothed density in $\mathbb{R}^{Md}$ is referred to as *M-density*:

$$p(\mathbf{y}) = \int p(\mathbf{y}|x)p(x)dx, \text{ where } p(\mathbf{y}|x) = \prod_{m=1}^{M} p_m(y_m|x) \quad (1)$$

- Note the convention $p(\mathbf{y}) := p_{\mathbf{Y}}(\mathbf{y}),\; p(x) := p_X(x)$, etc.
- Given $\{x_i\}_{i=1}^n$, how can we learn $p(\mathbf{y})$? Learn how to estimate X given $\mathbf{Y} = \mathbf{y}$. We generalized empirical Bayes and derived analytical expression(s) for $\widehat{x}(\mathbf{y})$.
- How do we sample X? Sample $\mathbf{Y}$ with Langevin MCMC and estimate X.

Parametrization Schemes



(a) x (b) $\mathbf{y} = (y_1, \ldots, y_M)$ (c) $\widehat{x}_\theta^{(m)}(\mathbf{y})$ (d) $\bar{\mathbf{y}}$

- MDAE parametrization scheme:

$$\nabla log\ p(\mathbf{y}) \approx \frac{\boldsymbol{\nu}_\theta(\mathbf{y}) - \mathbf{y}}{\sigma^2}, \tag{2}$$

  where $\boldsymbol{\nu}$ is a neural network and σ is the noise level: $Y_m = X + N(0, \sigma^2 I_d)$
- MDAE learning objective:

$$\mathcal{L}(\theta) = \frac{1}{M} \mathbb{E}_{(x,\mathbf{y}) \sim p(\mathbf{y}|x)p(x)} \left\| x \otimes M - \boldsymbol{\nu}_\theta(\mathbf{y}) \right\|_2^2, \text{ where } x \otimes M := (x, x, \ldots, x) \tag{3}$$

- Note that we do *not* use MCMC during learning.
- The first theoretical connection between DAEs and empirical Bayes.

Walk-Jump Sampling (Saremi and Hyvärinen, 2019)

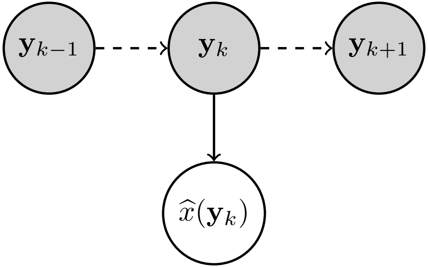


Figure: walk-jump sampling

- We use **underdamped** Langevin MCMC to sample the learned M-density $p_\theta(\mathbf{y})$.
- MCMC walk is based on discretizing the underdamped Langevin diffusion:

$$d\mathbf{v}_t = -\gamma \mathbf{v}_t dt + u \nabla \log p_\theta(\mathbf{y}_t) dt + (\sqrt{2\gamma u}) d\mathbf{B}_t,$$
$$d\mathbf{y}_t = \mathbf{v}_t dt.$$

- Jumps are decoupled from the MCMC chain:

$$\widehat{x}_\theta(\mathbf{y}) = \nu_\theta(\mathbf{y}) \tag{4}$$

- This sampling scheme is exact for large M.

Permutation-Invariant Gaussian M-Densities

- **Stable chains with 1 million steps on MNIST and CIFAR-10.**
- The first **unconditional** MCMC-based generative model on FFHQ-256.
- Note that sampling in diffusion models is based on a *sequence of conditional distributions*.
- Figure on the left: $M = 8$ and $\sigma = 4$.
- M-density is permutation invariant.
- $Md \approx 1.5 \times 10^6$
- Only 10 steps per image in the chain visualized here (starting from noise).
- Smoothing is key for good mixing.

FFHQ-256

Figure: Three MCMC chains for FFHQ-256 MDAE ($\sigma = 4,\ M = 8$)

CIFAR-10

- FID score of 43.95 obtained from a *single* MCMC chain (50,000 samples selected from a single chain with 1 million steps).
- State of the art in terms of the FID score obtained for long chains.
- By far the longest chain (1 million+ steps) in the literature.
- FID score 21.74 by increasing M and improved training (work in progress):