NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2023

BETA DIFFUSION

Mingyuan Zhou, Tianqi Chen, Zhendong Wang, Huangjie Zheng

DECEMBER 2023

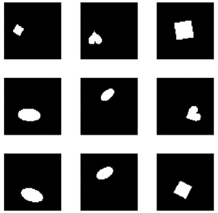
TEXAS
The University of Texas at Austin

The University of Texas at Austin
McCombs School of Business

The University of Texas at Austin
Department of Statistics
and Data Sciences

Is "Gaussian diffusion" all you need?

- Gaussian diffusion models excel in generating high-dimensional continuous data.
- To effectively address diverse data types, such as those marked by sparsity, skewness, heavy-tailedness, overdispersion, discreteness, and/or bounded ranges, our motivation lies in constructing new families of diffusion models:



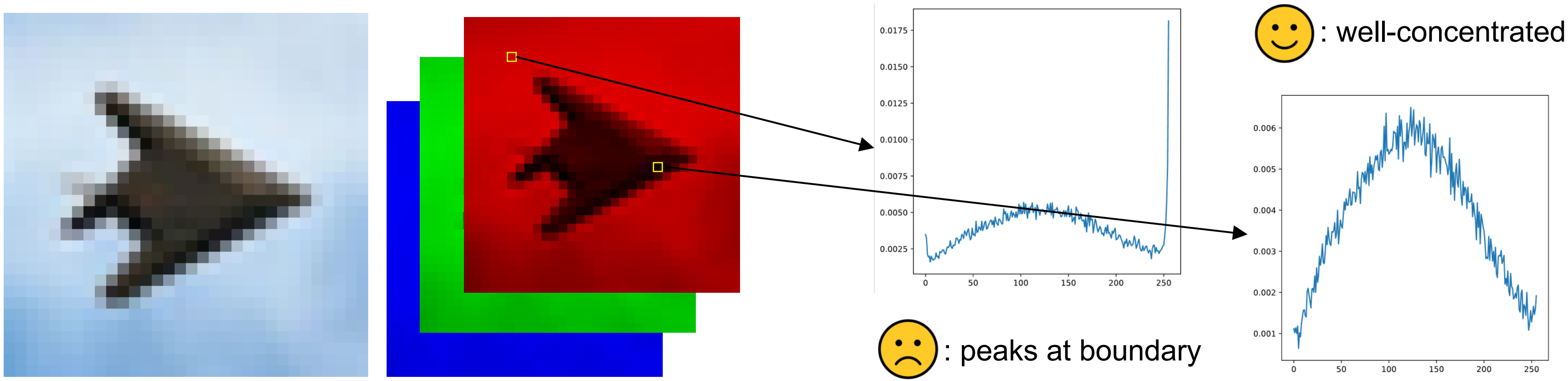
	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0

Left image source: https://www.tensorflow.org/datasets/catalog/dsprites / Right image source: https://hbctraining.github.io/scRNA-seq/lessons/02_SC_generation_of_count_matrix.html

- Exiting works:
  - Binary/categorical diffusion for binary/categorical data
  - Poisson diffusion (learning to jump) for count data
- Our work: Beta diffusion for range-bounded continuous data

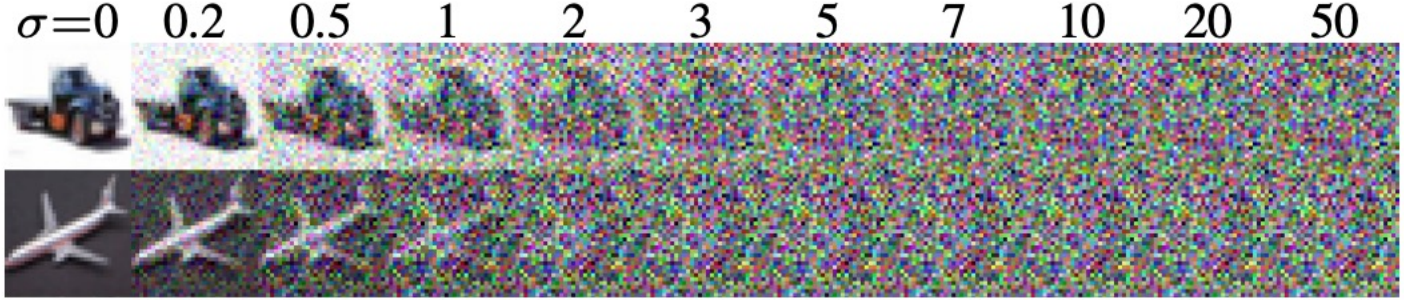
Range bounded data

- Blood pressure, oxygen level, age, body weight, height
- 8-bit image pixel values have boundaries at 0 and 255



Gaussian Diffusion (learning to denoise)

- Forward diffusion process (additive noise becomes larger and larger)



- Reverse diffusion process (denoising-based iterative refinement)

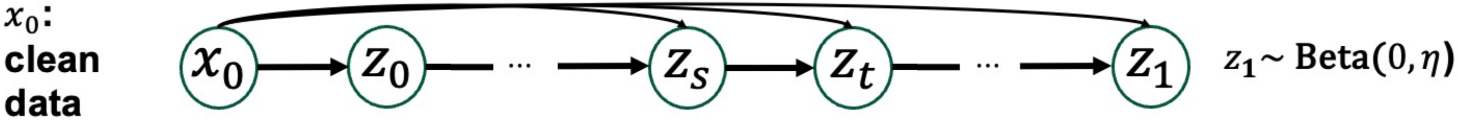
Image credit: Karras et al. (2022), Elucidating the Design Space of Diffusion-Based Generative Models

Beta forward diffusion

- Illustration:

Illustration of the beta forward diffusion process for two example images. The first column displays the original images, while the other 21 columns display the images noised and masked by beta diffusion.

- Forward diffusion chain (non-Markovian):



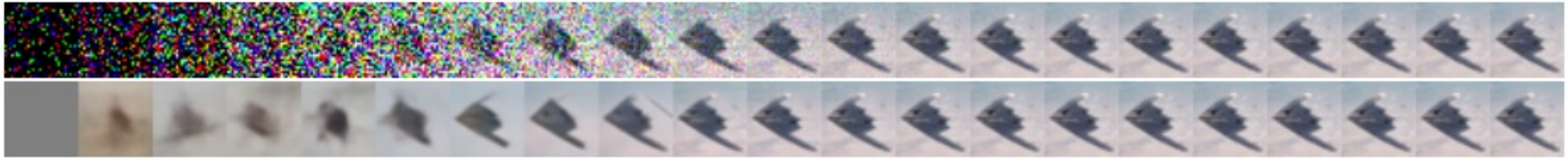
$$z_t = z_s \pi_{s\to t}, \quad \pi_{s\to t} \sim \text{Beta}(\eta\alpha_t x_0, \eta(\alpha_s - \alpha_t)x_0)$$

- Analytic forward marginal:

$$q(z_t \,|\, x_0) = \text{Beta}(\eta\alpha_t x_0, \eta(1-\alpha_t x_0))$$

Beta reverse diffusion

- Illustration:



- Reverse diffusion chain:

$$z_s = z_t + (1 - z_t)p_{s\leftarrow t}, \quad p_{s\leftarrow t} \sim \mathrm{Beta}(\eta(\alpha_s - \alpha_t)x_0, \eta(1 - \alpha_s x_0))$$

Generative modeling

$$z_s|z_t = z_t + (1 - z_t)\mathbf{Beta}\big(\eta(\alpha_s - \alpha_t)\hat{x}_0, \eta(1 - \alpha_t \hat{x}_0)\big), \qquad \hat{x}_0 = f_\theta(z_t, t)$$

- Analytic reverse conditional (scaled and shifted beta distribution):

$$q(z_s \mid z_t, x_0) = \frac{1}{1-z_t}\mathrm{Beta}\left(\frac{z_s - z_t}{1-z_t}; \eta(\alpha_s - \alpha_t)x_0, \eta(1 - \alpha_s x_0)\right)$$

Beta diffusion (training and generation)

- Injecting noise $q(z_t \mid x_0) = \mathbf{Beta}(\eta\alpha_t x_0, \eta(1-\alpha_t x_0))$
- Training via KL-divergence upper bounds (KLUBs):
  - KLUBs:
    $$\mathbf{KL}(q(z_s \mid z_t, \hat{x}_0 = f_\theta(z_t, t))||q(z_s \mid z_t, x_0))]$$
    $$\mathbf{KL}(q(z_t' \mid f_\theta(z_t, t))||q(z_t' \mid x_0))$$
  - Optimal solution:
    $$f_{\theta^*}(z_t, t) = \mathbb{E}[x_0 \mid z_t] = \mathbb{E}_{x_0 \sim q(x_0 \mid z_t)}[x_0]$$
- Generation: Sample $\hat{x}_0 = f_\theta(z_t, t)$ is refined iteratively

$$z_s|z_t = z_t + (1-z_t)\mathbf{Beta}\big(\eta(\alpha_s-\alpha_t)\hat{x}_0, \eta(1-\alpha_t\hat{x}_0)\big), \qquad \hat{x}_0 = f_\theta(z_t, t)$$

Beta diffusion and Bregman divergence

- KLUBs and Bregman divergence:

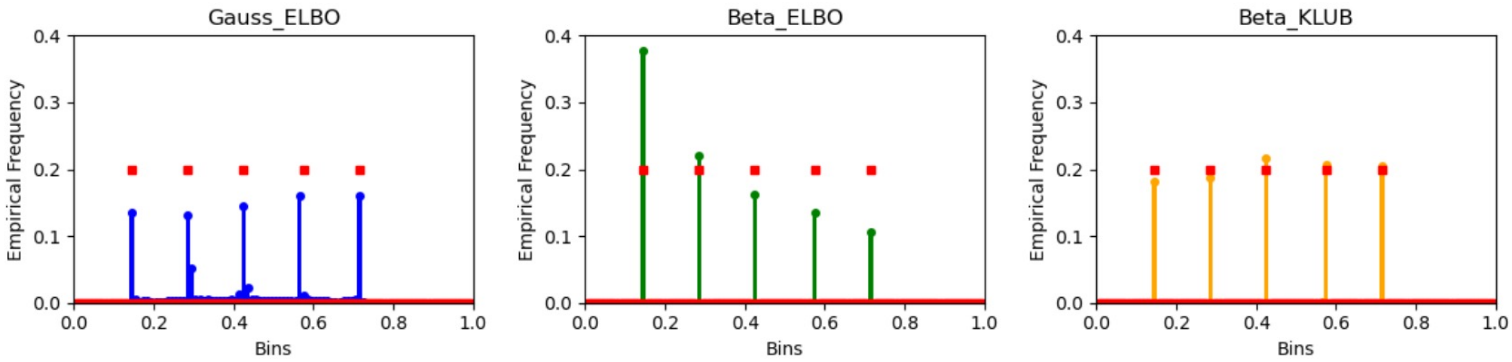
$$\begin{aligned}\mathrm{KL}(\mathrm{Beta}(\alpha_p, \beta_p)||\mathrm{Beta}(\alpha_q, \beta_q)) &= \ln \frac{B(\alpha_q, \beta_q)}{B(\alpha_p, \beta_p)} - (\alpha_q - \alpha_p, \beta_q - \beta_p) \begin{pmatrix} \nabla_\alpha \ln B(\alpha_p, \beta_p) \\ \nabla_\beta \ln B(\alpha_p, \beta_p) \end{pmatrix} \\ &= D_{\ln B(a,b)}((\alpha_q, \beta_q), (\alpha_p, \beta_p)).\end{aligned}$$

- Both KLUBs can be expressed as a Bregman divergence with x_0 in the 1st argument and f_θ in the 2nd argument.
- Therefore, the KLUBs are minimized at θ^* when

$$f_{\theta^*}(z_t, t) = \mathbb{E}[x_0 \mid z_t, t] \text{ for all } z_t \sim q(z_t)$$

Synthetic Data

- An equal mixture of five unit point masses (unit point mass can also be seen as an extreme case of range-bounded data, where the range is zero):



CIFAR10 images

- By adapting the CIFAR10 VP-EDM code, originally optimized for Gaussian diffusion, to our model, we can already achieve competitive results in FID

Table 2: Comparing FID scores for KLUB and negative ELBO-optimized Beta Diffusion on CIFAR-10 with varying NFE under $\eta = 10000$ and two different mini-batch sizes B.

Loss B	−ELBO 512	−ELBO 288	KLUB 512	KLUB 288
20	16.04	16.10	17.06	16.09
50	6.82	6.82	6.48	5.96
200	4.55	4.84	3.69	3.31
500	4.39	4.65	3.45	3.10
1000	4.41	4.61	3.38	3.08
2000	4.50	4.66	3.37	**3.06**

(a) −ELBO (b) KLUB

Figure 4: Uncurated randomly-generated images by beta diffusion optimized with −ELBO or KLUB with $\eta = 10000$ and $B = 288$. The generation with NFE = 1000 starts from the same random seed.

CIFAR10 images

- Quantitative evaluation

Diffusion Space	Model	FID (↓)
Gaussian	DDPM [23]	3.17
	VDM [35]	4.00
	Improved DDPM [46]	2.90
	TDPM+ [78]	2.83
	VP-EDM [34]	**1.97**
Gaussian+Blurring	Soft Diffusion [12]	3.86
	Blurring Diffusion [25]	3.17
Deterministic	Cold Diffusion (image reconstruction) [5]	80.08 (deblurring) 8.92 (inpainting)
	Inverse Heat Dispersion [52]	18.96
Categorical	D3PM Gauss+Logistic [2]	7.34
	τLDR-10 [8]	3.74
Count	JUMP (Poisson Diffusion) [10]	4.80
Range-bounded	Beta Diffusion	3.06

Acknowledgements

- Funding/Computing Support