Star-Shaped Denoising Diffusion Probabilistic Models

Andrey Okhotin*, Dmitry Molchanov*, Vladimir Arkhipkin,
Grigory Bartosh, Aibek Alanov, Dmitry P. Vetrov

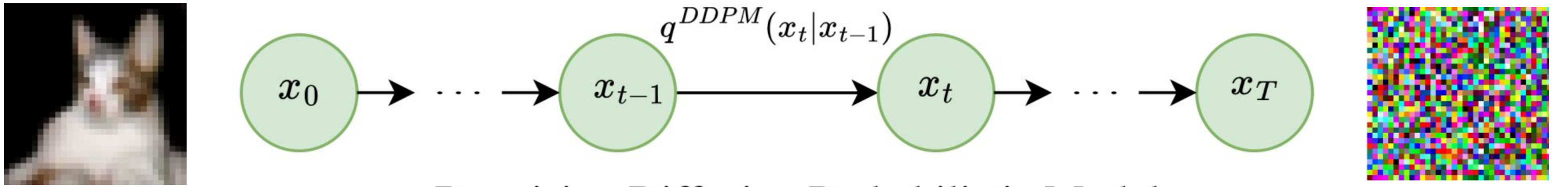
HSE University

DDPM

The Gaussian DDPM (Ho et al., 2020) is defined as a forward (diffusion) process and a corresponding reverse (denoising) process DDPM. The forward process is defined as a Markov chain with Gaussian conditionals:

$$q^{\mathrm{DDPM}}(x_{0:T}) = q(x_0)\prod_{t=1}^{T} q^{\mathrm{DDPM}}(x_t|x_{t-1}) \tag{1}$$

$$q^{\mathrm{DDPM}}(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}\right) \tag{2}$$



Denoising Diffusion Probabilistic Models

DDPM

Learnable reverse process follow a similar structure and constitutes a generative part of the model:

$$p_\theta^{\mathrm{DDPM}}(x_{0:T}) = q^{\mathrm{DDPM}}(x_T)\prod_{t=1}^{T} p_\theta^{\mathrm{DDPM}}(x_{t-1}|x_t) \quad (3)$$

$$p_\theta^{\mathrm{DDPM}}(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\right) \quad (4)$$

$$\mathcal{L}^{\mathrm{DDPM}}(\theta) = \mathbb{E}_{q^{\mathrm{DDPM}}}\Big[\log p_\theta^{\mathrm{DDPM}}(x_0|x_1) - \quad (5)$$

$$-\sum_{t=2}^{T} D_{KL}\left(q^{\mathrm{DDPM}}(x_{t-1}|x_t, x_0) \,\|\, p_\theta^{\mathrm{DDPM}}(x_{t-1}|x_t)\right)\Big] \quad (6)$$

$$\mathcal{L}^{\mathrm{DDPM}}(\theta) \to \max_\theta \quad (7)$$

Let's try to change noise in MC

$$x_0 \in \mathbb{M}, \quad x_t \sim q^{\mathrm{DDPM}}(x_t|x_0) \not\Rightarrow x_t \in \mathbb{M}$$

$$q^{\mathrm{DDPM}}(x_t|x_{t-1}) = \{\ x_{t-1} \in \mathbb{M} \Rightarrow x_t \in \mathbb{M}\ \}$$

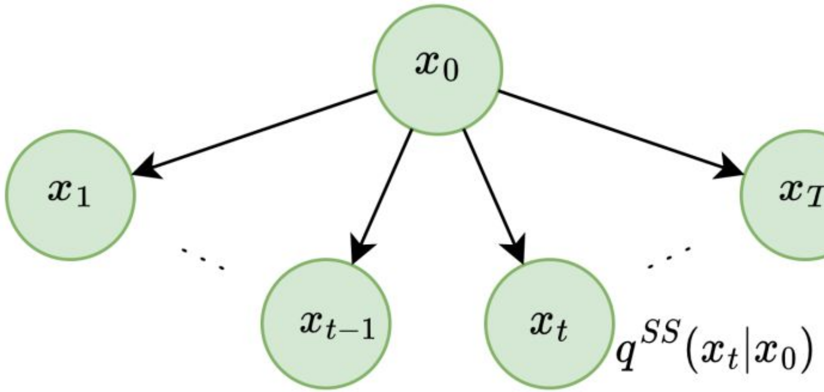
$$x_0 \in B_1(0): \quad q^{\mathrm{DDPM}}(x_t|x_{t-1}) \in vMF\,(\ldots) \Rightarrow x_t \in B_1(0)$$

$$q^{\mathrm{DDPM}}(x_{t-1}|x_t, x_0) = \frac{q^{\mathrm{DDPM}}(x_t|x_{t-1})q^{\mathrm{DDPM}}(x_{t-1}|x_0)}{q^{\mathrm{DDPM}}(x_t|x_0)}$$

no analytical form of $q^{\mathrm{DDPM}}(x_t|x_0)$ for most $q^{\mathrm{DDPM}}(x_t|x_{t-1})$

Star-Shaped (SS-DDPM)

$$q^{\mathrm{SS}}(x_{0:T}) = q(x_0)\prod_{t=1}^{T} q^{\mathrm{SS}}(x_t|x_0), \qquad (8)$$



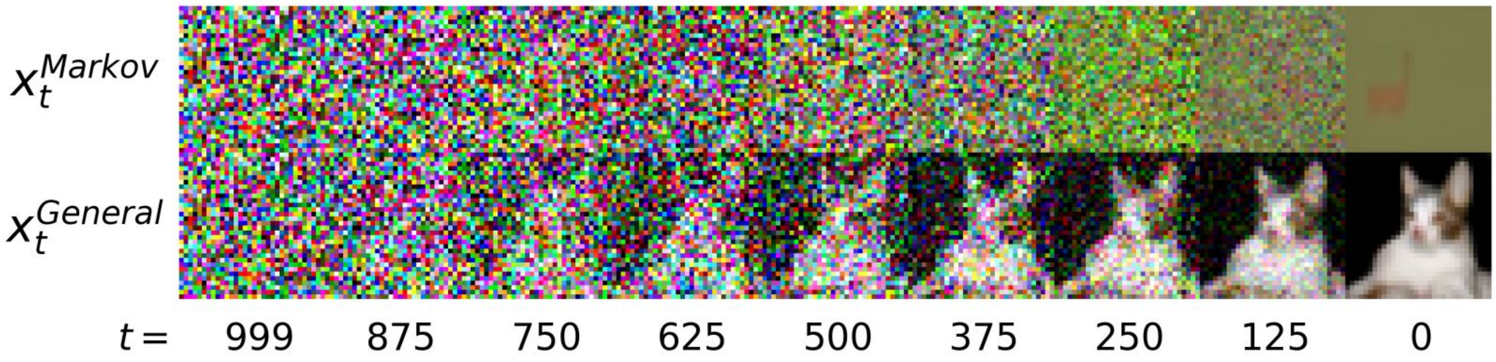
Star-Shaped Denoising Diffusion Probabilistic Models

Star-Shaped

$$q^{\mathrm{SS}}(x_{0:T}) = q(x_0)\prod_{t=1}^{T} q^{\mathrm{SS}}(x_t|x_0), \tag{8}$$

$$q^{\mathrm{DDPM}}(x_{0:T}) = q^{\mathrm{DDPM}}(x_T)\prod_{t=1}^{T} q^{\mathrm{DDPM}}(x_{t-1}|x_t). \tag{9}$$

$$q^{\mathrm{SS}}(x_{0:T}) = q^{\mathrm{SS}}(x_T)\prod_{t=1}^{T} q^{\mathrm{SS}}(x_{t-1}|x_{t:T}) \tag{10}$$



Star-Shaped VLB

$$\mathcal{L}^{\text{SS}}(\theta) = \mathbb{E}_{q^{\text{SS}}(x_{0:T})} \log \frac{p_\theta^{\text{SS}}(x_{0:T})}{q^{\text{SS}}(x_{1:T}|x_0)} = \mathbb{E}_{q^{\text{SS}}(x_{0:T})} \log \frac{p_\theta^{\text{SS}}(x_0|x_{1:T}) p_\theta^{\text{SS}}(x_T) \prod_{t=2}^{T} p_\theta^{\text{SS}}(x_{t-1}|x_{t:T})}{\prod_{t=1}^{T} q^{\text{SS}}(x_t|x_0)} = \tag{26}$$

$$= \mathbb{E}_{q^{\text{SS}}(x_{0:T})} \left[\log p_\theta^{\text{SS}}(x_0|x_{1:T}) + \sum_{t=2}^{T} \log \frac{p_\theta^{\text{SS}}(x_{t-1}|x_{t:T})}{q^{\text{SS}}(x_{t-1}|x_0)} + \cancel{\log \frac{p_\theta^{\text{SS}}(x_T)}{q^{\text{SS}}(x_T|x_0)}} \right] = \tag{27}$$

$$= \mathbb{E}_{q^{\text{SS}}(x_{0:T})} \left[\log p_\theta^{\text{SS}}(x_0|x_{1:T}) - \sum_{t=2}^{T} D_{KL}\left(q^{\text{SS}}(x_{t-1}|x_0) \,\|\, p_\theta^{\text{SS}}(x_{t-1}|x_{t:T})\right) \right] \tag{28}$$

Star-Shaped

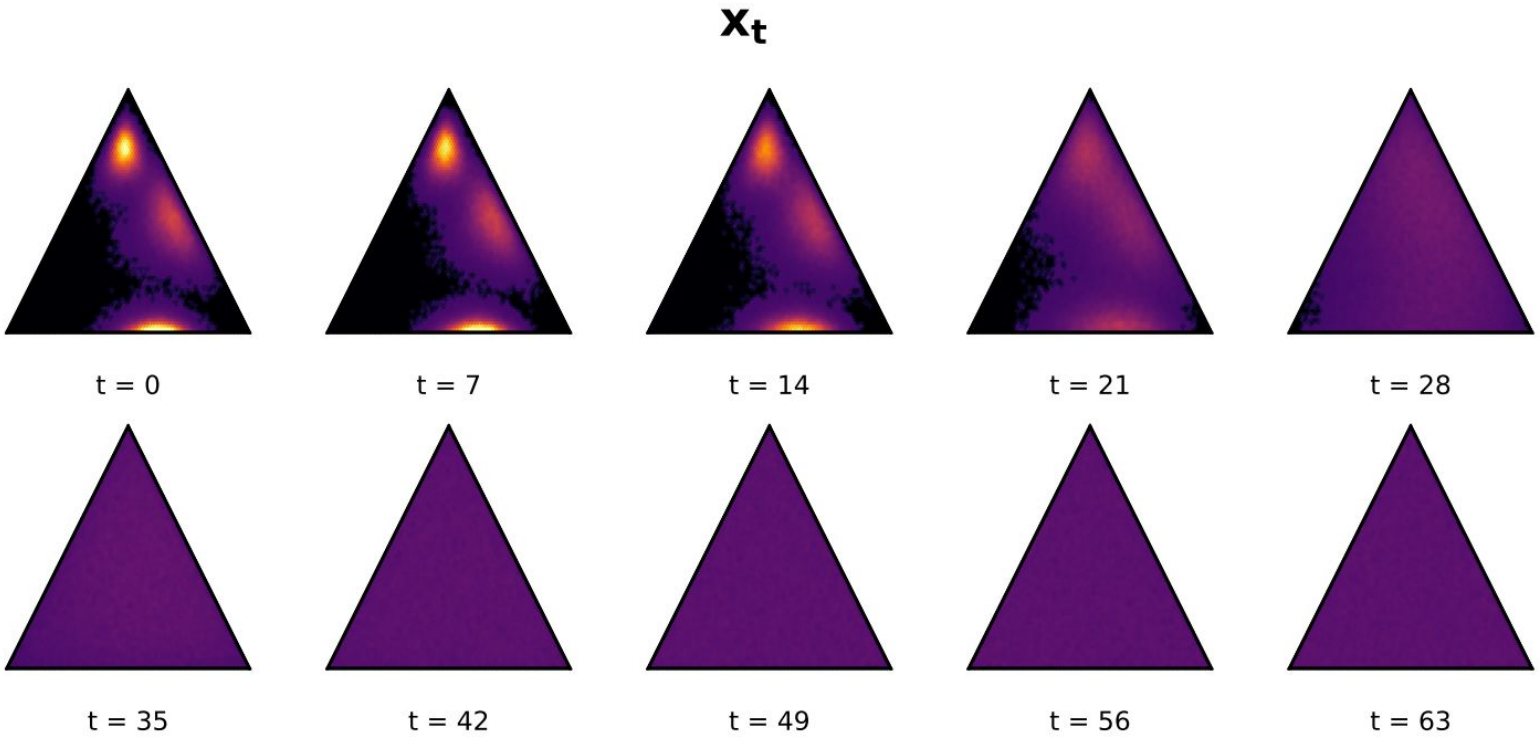
$$q^{\mathrm{SS}}(x_{0:T}) = q(x_0)\prod_{t=1}^{T} q^{\mathrm{SS}}(x_t|x_0), \tag{8}$$

$$q^{\mathrm{SS}}(x_{0:T}) = q^{\mathrm{SS}}(x_T)\prod_{t=1}^{T} q^{\mathrm{SS}}(x_{t-1}|x_{t:T}) \tag{10}$$

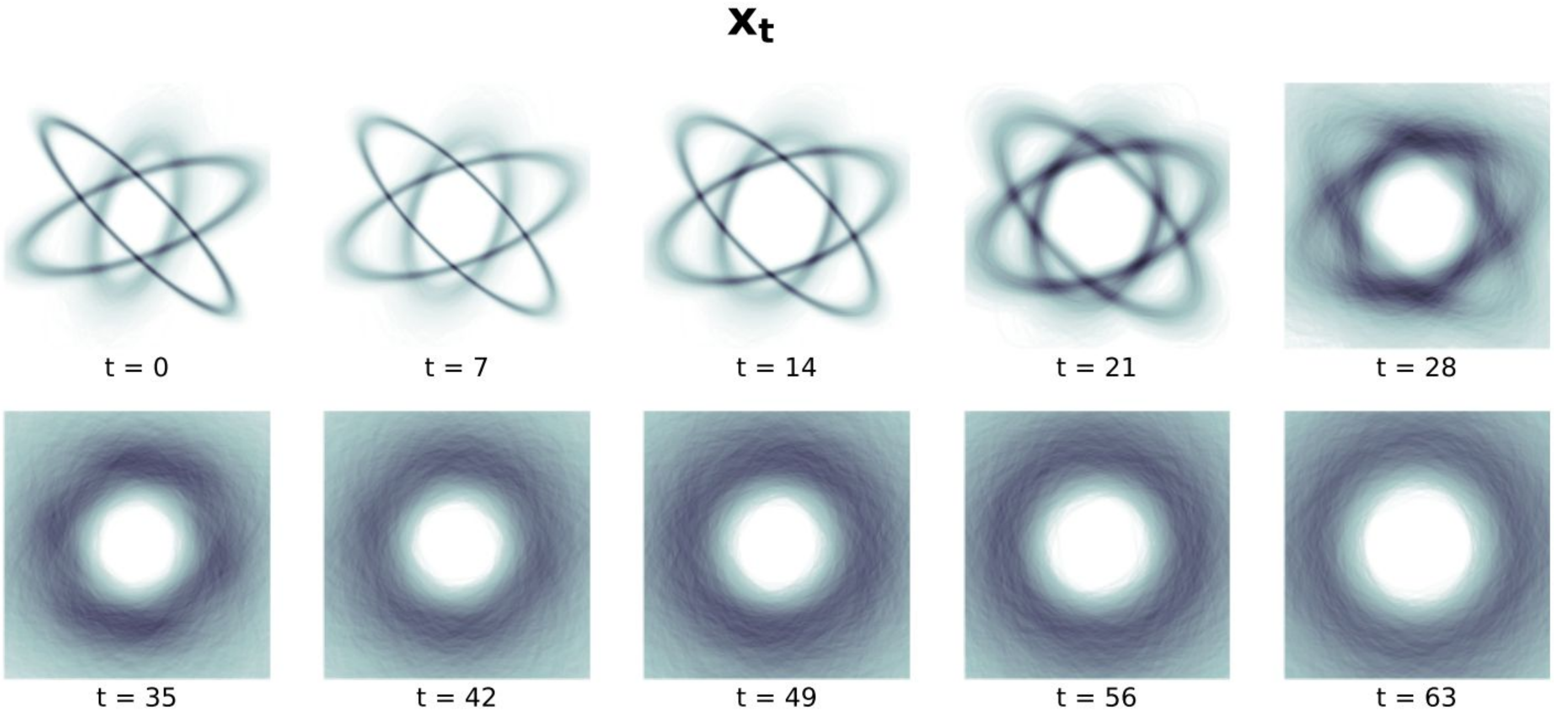
$$p_\theta^{\mathrm{SS}}(x_{0:T}) = p_\theta^{\mathrm{SS}}(x_T)\prod_{t=1}^{T} p_\theta^{\mathrm{SS}}(x_{t-1}|x_{t:T}) \tag{11}$$

$$\begin{aligned}\mathcal{L}^{\mathrm{SS}}(\theta) = \mathbb{E}_{q^{\mathrm{SS}}} \Big[&\log p_\theta(x_0|x_{1:T}) - \\ &- \sum_{t=2}^{T} D_{KL}\left(q^{\mathrm{SS}}(x_{t-1}|x_0) \,\|\, p_\theta^{\mathrm{SS}}(x_{t-1}|x_{t:T})\right)\Big]\end{aligned} \tag{12}$$

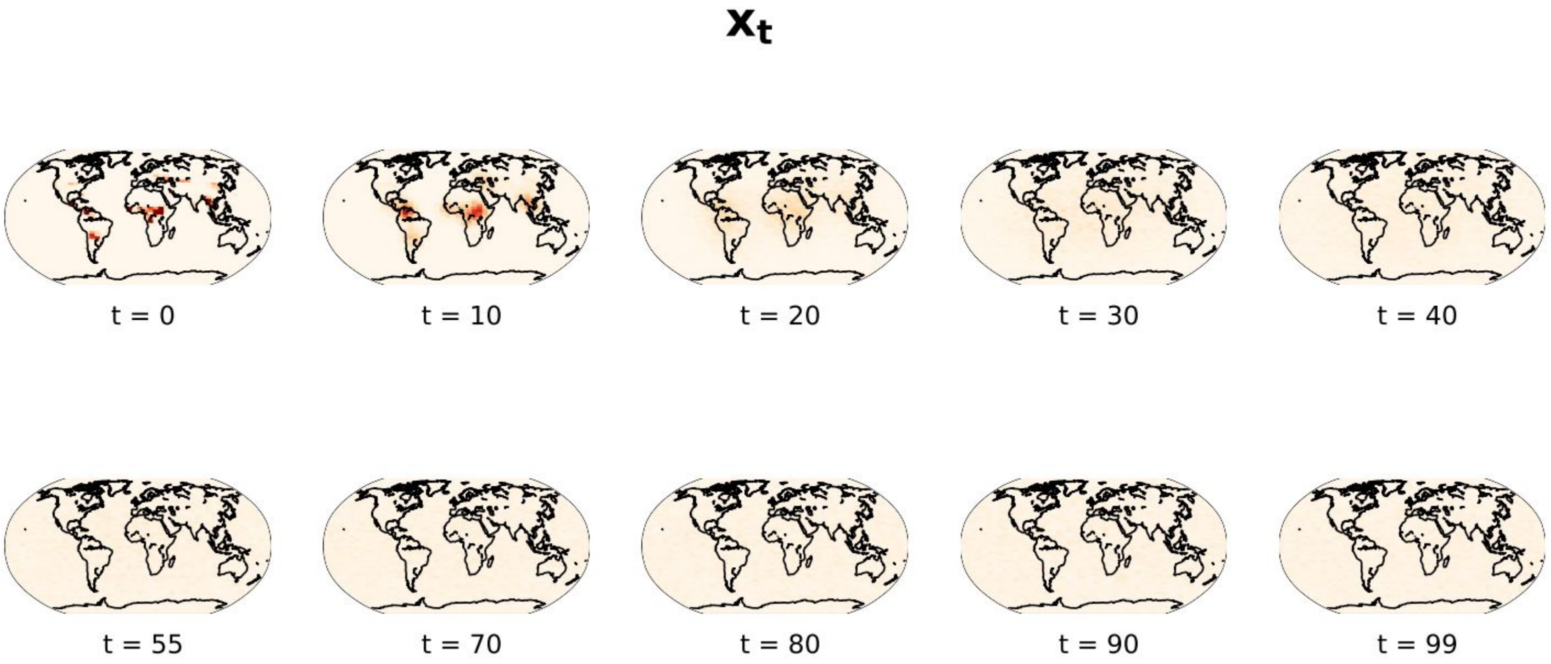
Dirichlet



Wishart



von Mises-Fisher



How to condition on the whole tail?

Approches:

How to condition on the whole tail?

Approches:

1. **easy**: concatenate all objects from tail and use them as an input to NN

How to condition on the whole tail?

Approches:

1. **easy**: concatenate all objects from tail and use them as an input to NN
2. **medium**: use LSTM-like NN architecture

How to condition on the whole tail?

Approches:

1. **easy**: concatenate all objects from tail and use them as an input to NN
2. **medium**: use LSTM-like NN architecture
3. **advanced**: compress information from the whole tail to one object with fixed size

$$q^{\mathrm{SS}}(x_{t-1}|x_{t:T}) = q^{\mathrm{SS}}(x_{t-1}|G_t) \qquad (13)$$

Efficient tail conditioning

Theorem 1. *Given*

$$q^{\mathrm{SS}}(x_t|x_0) = h_t(x_t)\exp\left\{\eta_t(x_0)^{\mathsf{T}}\mathcal{T}(x_t) - \Omega_t(x_0)\right\} \tag{14}$$

$$\eta_t(x_0) = A_t f(x_0) + b_t \tag{15}$$

$$G_t = \mathcal{G}_t(x_{t:T}) = \textstyle\sum_{s=t}^{T} A_s^{\mathsf{T}}\mathcal{T}(x_s) \tag{16}$$

the following holds:

$$q^{\mathrm{SS}}(x_{t-1}|x_{t:T}) = q^{\mathrm{SS}}(x_{t-1}|G_t) \tag{17}$$

Proof.

$$\boxed{q^{\mathrm{SS}}(x_t|x_0) = h_t(x_t)\exp\left\{\eta_t(x_0)^\mathsf{T}\mathcal{T}(x_t) - \Omega_t(x_0)\right\}}$$

$$\boxed{\eta_t(x_0) = A_t f(x_0) + b_t}$$

$$q^{\mathrm{SS}}(x_{t-1}|x_{t:T}) = \int q^{\mathrm{SS}}(x_{t-1}|x_0)q^{\mathrm{SS}}(x_0|x_{t:T})dx_0 \tag{40}$$

$$q^{\mathrm{SS}}(x_0|x_{t:T}) = \frac{q(x_0)\prod_{s=t}^{T} q^{\mathrm{SS}}(x_s|x_0)}{q^{\mathrm{SS}}(x_{t:T})} = \frac{q(x_0)}{q^{\mathrm{SS}}(x_{t:T})}\left(\prod_{s=t}^{T} h_s(x_s)\right)\exp\left\{\sum_{s=t}^{T}\left(\eta_s(x_0)^\mathsf{T}\mathcal{T}(x_s) - \Omega_s(x_0)\right)\right\} = \tag{41}$$

$$= \frac{q(x_0)}{q^{\mathrm{SS}}(x_{t:T})}\left(\prod_{s=t}^{T} h_s(x_s)\right)\exp\left\{\sum_{s=t}^{T}\left((A_s f(x_0) + b_s)^\mathsf{T}\mathcal{T}(x_s) - \Omega_s(x_0)\right)\right\} = \tag{42}$$

$$= \frac{q(x_0)}{q^{\mathrm{SS}}(x_{t:T})}\left(\prod_{s=t}^{T} h_s(x_s)\right)\exp\left\{f(x_0)^\mathsf{T}\sum_{s=t}^{T} A_s^\mathsf{T}\mathcal{T}(x_s) + \sum_{s=t}^{T}\left(b_s^\mathsf{T}\mathcal{T}(x_s) - \Omega_s(x_0)\right)\right\} = \tag{43}$$

$$= \frac{q(x_0)}{q^{\mathrm{SS}}(x_{t:T})}\left(\prod_{s=t}^{T} h_s(x_s)\right)\exp\left\{f(x_0)^\mathsf{T}G_t + \sum_{s=t}^{T}\left(b_s^\mathsf{T}\mathcal{T}(x_s) - \Omega_s(x_0)\right)\right\} = \tag{44}$$

$$= \frac{q(x_0)\exp\left\{f(x_0)^\mathsf{T}G_t - \sum_{s=t}^{T}\Omega_s(x_0)\right\}}{\int q(x_0)\exp\left\{f(x_0)^\mathsf{T}G_t - \sum_{s=t}^{T}\Omega_s(x_0)\right\}dG_t} = q^{\mathrm{SS}}(x_0|G_t) \tag{45}$$

$$q^{\mathrm{SS}}(x_{t-1}|x_{t:T}) = \int q^{\mathrm{SS}}(x_{t-1}|x_0)q^{\mathrm{SS}}(x_0|x_{t:T})dx_0 = \int q^{\mathrm{SS}}(x_{t-1}|x_0)q^{\mathrm{SS}}(x_0|G_t)dx_0 = q^{\mathrm{SS}}(x_{t-1}|G_t) \tag{46}$$

Training

$$p_\theta^{\mathrm{SS}}(x_{t-1}|x_{t:T}) \approx q^{\mathrm{SS}}(x_{t-1}|x_{t:T}) = \\ = \int q^{\mathrm{SS}}(x_{t-1}|x_0)q^{\mathrm{SS}}(x_0|x_{t:T})dx_0 \tag{20}$$

Algorithm 1 SS-DDPM training

repeat

$x_0 \sim q(x_0)$

$t \sim \mathrm{Uniform}(1, \ldots, T)$

$x_{t:T} \sim q^{\mathrm{SS}}(x_{t:T}|x_0)$

$G_t = \sum_{s=t}^{T} A_s^{\mathsf{T}} \mathcal{T}(x_s)$

Move along $\nabla_\theta \mathrm{KL}(q^{\mathrm{SS}}(x_{t-1}|x_0) \| p_\theta^{\mathrm{SS}}(x_{t-1}|G_t))$

until Convergence

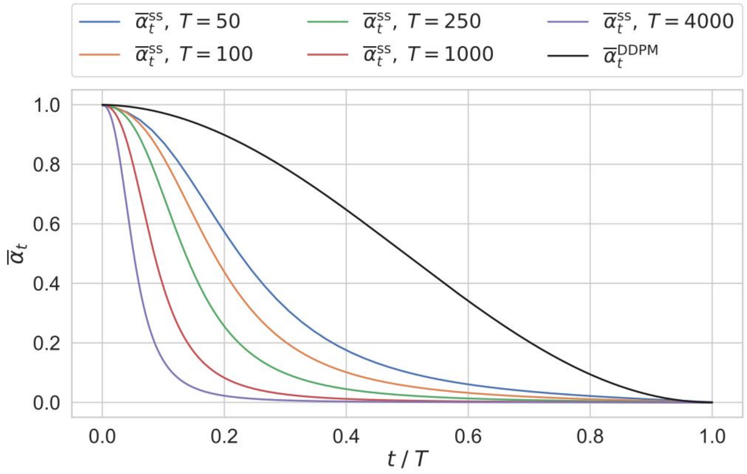
Sampling

$$p_\theta^{\text{SS}}(x_{t-1}|x_{t:T}) = q^{\text{SS}}(x_{t-1}|x_0)\big|_{x_0 = x_\theta(\mathcal{G}_t(x_{t:T}),t)} \quad (21)$$

Algorithm 2 SS-DDPM sampling

$x_T \sim q^{\text{SS}}(x_T)$
$G_T = A_T^\mathsf{T}\mathcal{T}(x_T)$
for $t = T$ to 2 **do**
  $\tilde{x}_0 = x_\theta(G_t, t)$
  $x_{t-1} \sim q^{\text{SS}}(x_{t-1}|x_0)\big|_{x_0=\tilde{x}_0}$
  $G_{t-1} = G_t + A_{t-1}^\mathsf{T}\mathcal{T}(x_{t-1})$
end for
$x_0 \sim p_\theta^{\text{SS}}(x_0|G_1)$

Connection with DDPM



Theorem 2. *Let $\overline{\alpha}_t^{\mathrm{DDPM}}$ define the noising schedule for a DDPM model (1–2) via $\beta_t = (\overline{\alpha}_{t-1}^{\mathrm{DDPM}} - \overline{\alpha}_t^{\mathrm{DDPM}})/\overline{\alpha}_{t-1}^{\mathrm{DDPM}}$. Let $q^{\mathrm{SS}}(x_{0:T})$ be a Gaussian SS-DDPM forward process with the following noising schedule and tail statistic:*

$$q^{\mathrm{SS}}(x_t|x_0) = \mathcal{N}\left(x_t; \sqrt{\overline{\alpha}_t^{\mathrm{SS}}}x_0, 1 - \overline{\alpha}_t^{\mathrm{SS}}\right), \tag{22}$$

$$\mathcal{G}_t(x_{t:T}) = \frac{1 - \overline{\alpha}_t^{\mathrm{DDPM}}}{\sqrt{\overline{\alpha}_t^{\mathrm{DDPM}}}} \sum_{s=t}^{T} \frac{\sqrt{\overline{\alpha}_s^{\mathrm{SS}}}x_s}{1 - \overline{\alpha}_s^{\mathrm{SS}}}, \text{ where} \tag{23}$$

$$\frac{\overline{\alpha}_t^{\mathrm{SS}}}{1 - \overline{\alpha}_t^{\mathrm{SS}}} = \frac{\overline{\alpha}_t^{\mathrm{DDPM}}}{1 - \overline{\alpha}_t^{\mathrm{DDPM}}} - \frac{\overline{\alpha}_{t+1}^{\mathrm{DDPM}}}{1 - \overline{\alpha}_{t+1}^{\mathrm{DDPM}}}. \tag{24}$$

Then the tail statistic G_t follows a Gaussian DDPM noising process $q^{\mathrm{DDPM}}(x_{0:T})|_{x_{1:T}=G_{1:T}}$ defined by the schedule $\overline{\alpha}_t^{\mathrm{DDPM}}$. Moreover, the corresponding reverse processes and VLB objectives are also equivalent.

Beta SS-DDPM

$$q(x_t|x_0) \;=\; \mathrm{Beta}(x_t; \alpha_t, \beta_t)$$

$$\alpha_t = 1 + \nu_t x_0$$

$$\beta_t = 1 + \nu_t (1 - x_0)$$

$$\eta_t(x_0) = \nu_t x_0, \mathcal{T}(x_t) = \log \tfrac{x_t}{1 - x_t}$$

$$\mathcal{G}_t(x_{t:T}) = \sum_{s=t}^{T} \nu_s \log \frac{x_s}{1 - x_s}$$

Theorem 1. *Given*

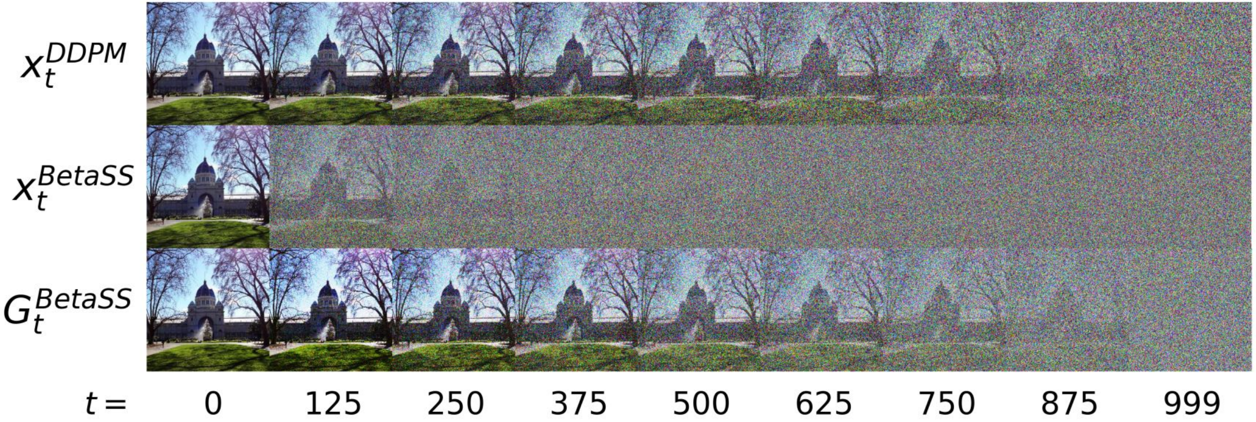
$$q^{\mathrm{SS}}(x_t|x_0) = h_t(x_t) \exp\left\{\eta_t(x_0)^{\mathsf{T}} \mathcal{T}(x_t) - \Omega_t(x_0)\right\} \quad (14)$$

$$\eta_t(x_0) = A_t f(x_0) + b_t \quad (15)$$

$$G_t = \mathcal{G}_t(x_{t:T}) = \textstyle\sum_{s=t}^{T} A_s^{\mathsf{T}} \mathcal{T}(x_s) \quad (16)$$

the following holds:

$$q^{\mathrm{SS}}(x_{t-1}|x_{t:T}) = q^{\mathrm{SS}}(x_{t-1}|G_t) \quad (17)$$



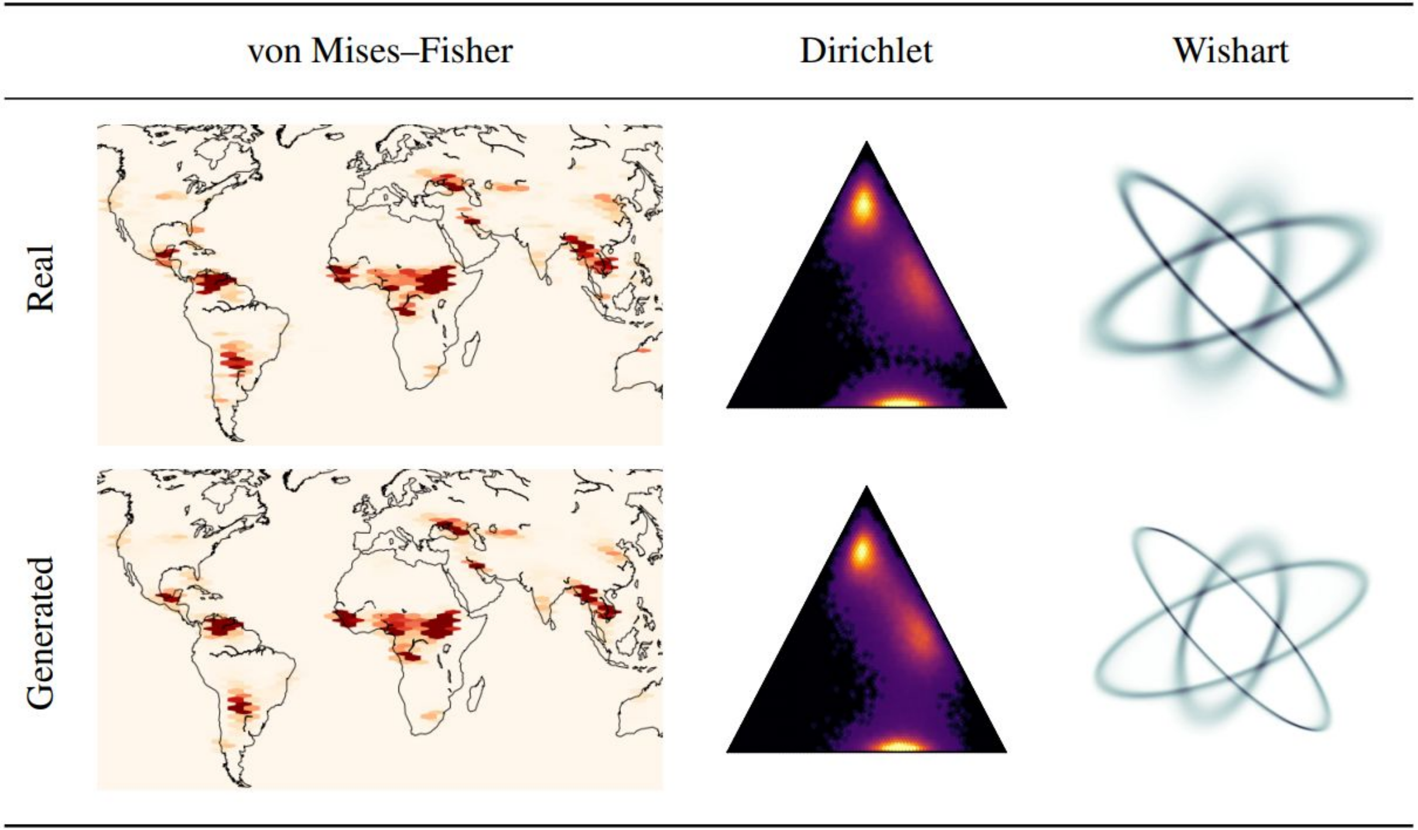
Practical considerations

1. choosing the right schedule
2. implementing the sampler
3. reducing the number of steps

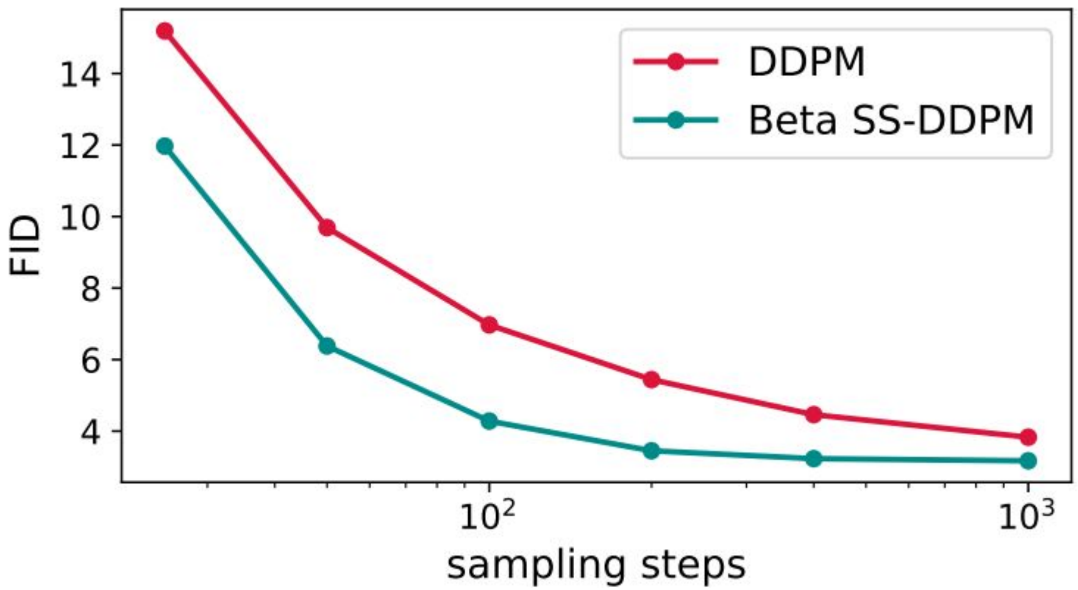
$$p_\theta^{\mathrm{SS}}(x_{t_1:t_2}|G_{t_2}) = \prod_{t=t_1}^{t_2} q^{\mathrm{SS}}(x_t|x_0)\big|_{x_0=x_\theta(G_t,t)} \approx \\ \approx \prod_{t=t_1}^{t_2} q^{\mathrm{SS}}(x_t|x_0)\big|_{x_0=x_\theta(G_{t_2},t_2)} \tag{25}$$

4. time-dependent tail normalization
5. architectural choices

Experiments



Beta SS-DDPM vs DDPM



Conclusion

SS-DDPM provided

1. approach for creating diffusion models with non-Gaussian noise
2. effective approach for Gaussian, Beta, Dirichlet, Categorical, von Mises, von Mises-Fisher, Gamma, Wishart distributions