From Denoising Diffusions to Denoising Markov Models

Joe Benton

University of Warwick
Friday 10th November 2023

Motivation

Diffusion Models

Figure: Images generated by DDPM [1], DALLE-2 [2] and Imagen [3].

Generative Modeling

The problem

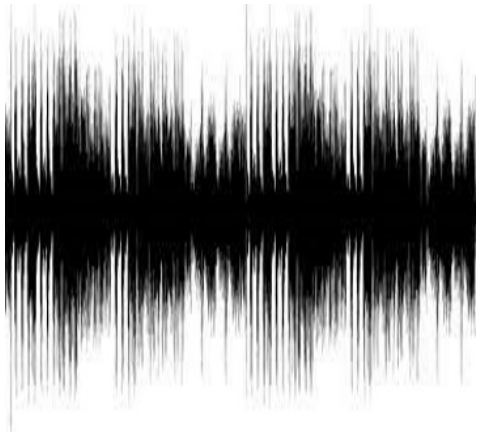
Given samples from a data distribution $p_{\text{data}}(\mathbf{x})$, generate synthetic samples coming from approximately the same distribution.

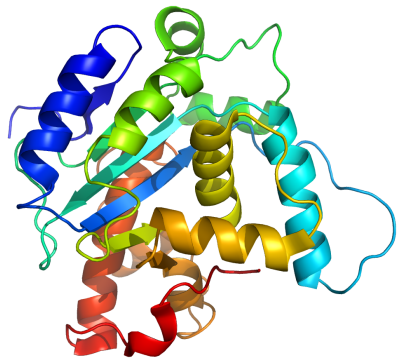
Generative Modeling

The problem

Given samples from a data distribution $p_{\text{data}}(\mathbf{x})$, generate synthetic samples coming from approximately the same distribution.

Applications: Image generation, text-to-speech, protein structure modeling, approximate posterior inference etc.





Motivation

But... these diffusion models are either restricted to data on $\mathbb{R}^d$, or rely on ad-hoc extensions to new state spaces.

Motivation

But... these diffusion models are either restricted to data on $\mathbb{R}^d$, or rely on ad-hoc extensions to new state spaces.

Motivating question

Can we find a principled generalisation of diffusion models to new state spaces?

Motivation

But... these diffusion models are either restricted to data on $\mathbb{R}^d$, or rely on ad-hoc extensions to new state spaces.

Motivating question

Can we find a principled generalisation of diffusion models to new state spaces?

Yes – Denoising Markov Models!

Brief Introduction to Diffusion Models

Diffusion models on $\mathbb{R}^d$

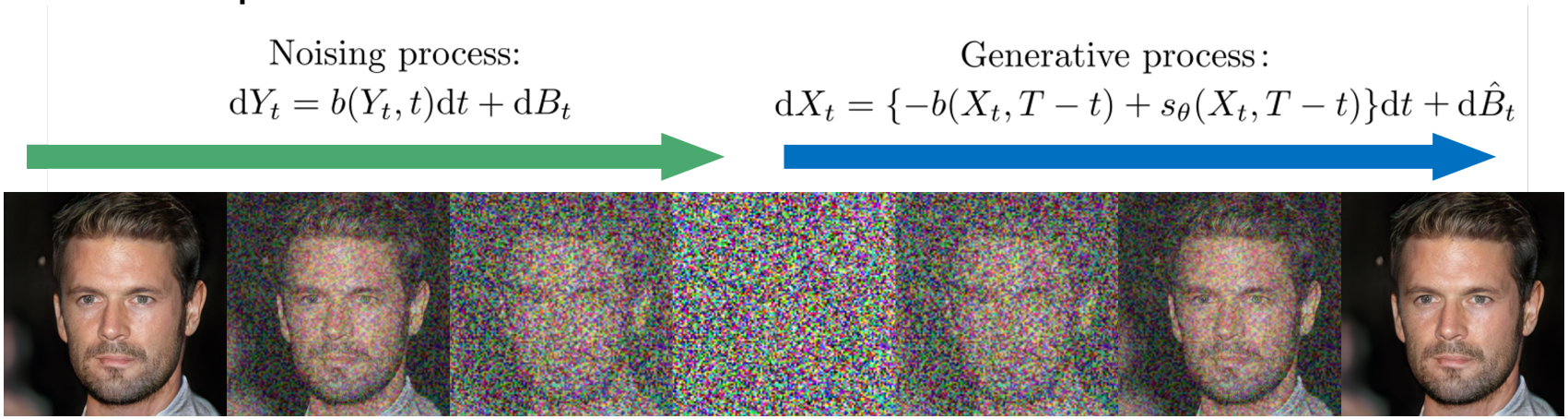
- Noising process $(Y_t)_{t\in[0,T]}$ with maringals $q_t(\mathbf{x})$ via the SDE

$$\mathrm{d}Y_t = -\tfrac{1}{2}Y_t\mathrm{d}t + \mathrm{d}B_t, \qquad Y_0 = \mathbf{x}_0 \sim p_{\text{data}}.$$

- Time-reversed process $X_t = Y_{T-t}$ satisfies

$$\mathrm{d}X_t = \{-\tfrac{1}{2}X_t + \nabla \log q_{T-t}(X_t)\}\mathrm{d}t + \mathrm{d}\hat{B}_t.$$

- **Strategy:** Learn approximation to $\nabla \log q_t(\mathbf{x})$, use to simulate reverse process.



Diffusion models on $\mathbb{R}^d$

- We approximate $\nabla \log q_t(\mathbf{x})$ using the L^2 objective

$$\mathcal{I}_{\text{DSM}}(\theta) = \frac{1}{2}\int_0^T \mathbb{E}_{q_{0,t}(\mathbf{x}_0,\mathbf{x}_t)}\left[||\nabla_{\mathbf{x}} \log q_{t|0}(\mathbf{x}_t|\mathbf{x}_0) - s_\theta(\mathbf{x}_t, t)||^2\right] \, \mathrm{d}t.$$

- $s_\theta(\mathbf{x}_t, t)$ is an approximation parameterised by a neural network.
- Originally proposed ad hoc; later derived by Huang et al. [4].

Score Matching

- A method for fitting unnormalized probability distributions of Hyvärinen [5].
- Approximate the distribution q_0 using parametric family $p(\mathbf{x};\theta) = q(x;\theta)/Z(\theta)$ by minimising

$$\mathcal{J}_{\text{ESM}}(\theta) = \frac{1}{2}\mathbb{E}_{q_0(\mathbf{x})}\left[||\nabla_{\mathbf{x}} \log q_0(\mathbf{x}) - \nabla_{\mathbf{x}} \log q(\mathbf{x};\theta)||^2\right].$$

- This is intractable, but equivalent to minimising

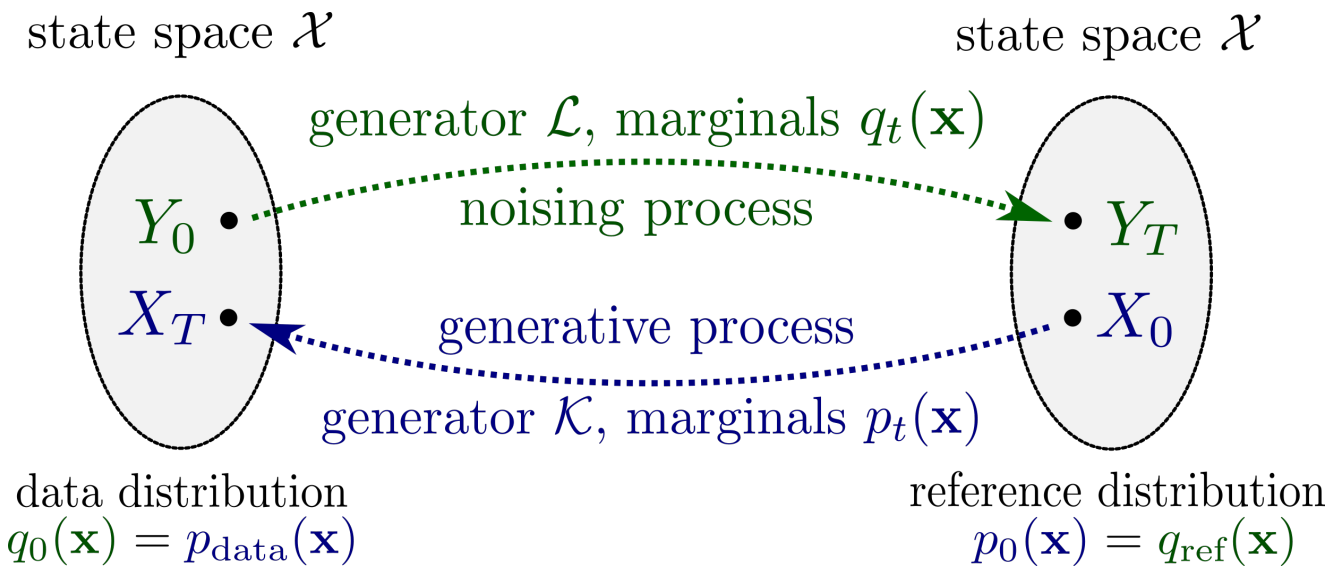
$$\mathcal{J}_{\text{ISM}}(\theta) = \mathbb{E}_{q_0(\mathbf{x})}\left[\Delta_{\mathbf{x}} \log q(\mathbf{x};\theta) + \frac{1}{2}\|\nabla_{\mathbf{x}} \log q(\mathbf{x};\theta)||^2\right],$$

or a denoising score matching objective.

Our Novel Framework: Denoising Markov Models

Denoising Markov Models

- $p_{\text{data}}(\mathbf{x})$ on space $\mathcal{X}$.
- Noising Markov process $(Y_t)_{t\in[0,T]}$, generator $\mathcal{L}$, marginals $q_t(\mathbf{x})$.
- Learn reverse process $(X_t)_{t\in[0,T]}$, generator $\mathcal{K}$, marginals $p_t(\mathbf{x})$.



Example

Euclidan Diffusion

If $(X_t)_{t\in[0,T]}$, $(Y_t)_{t\in[0,T]}$ are given by the SDEs

$$\begin{aligned} \mathrm{d}X_t &= \mu(X_t, t)\mathrm{d}t + \mathrm{d}\hat{B}_t, \\ \mathrm{d}Y_t &= b(Y_t, t)\mathrm{d}t + \mathrm{d}B_t, \end{aligned}$$

then the corresponding generators are

$$\begin{aligned} \mathcal{K} &= \partial_t + \mu \cdot \nabla + \tfrac{1}{2}\Delta, \\ \mathcal{L} &= \partial_t + b \cdot \nabla + \tfrac{1}{2}\Delta. \end{aligned}$$

Plan

Key question

How do we learn the reverse process generator $\mathcal{K}$?

Plan

Key question

How do we learn the reverse process generator $\mathcal{K}$?

The plan:

1. Model likelihood using Fokker–Planck, Feynman–Kac.
2. Lower bound on model log likelihood using Girsanov.
3. Equivalent tractable objectives.

Model Likelihood

(Generalised) Fokker–Planck PDE

$$\partial_t p_t = \hat{\mathcal{K}}^* p_t$$

Model Likelihood

(Generalised) Fokker–Planck PDE

$$\partial_t p_t = \hat{\mathcal{K}}^* p_t$$

Assumption 1

With $v(\mathbf{x}, t) = p_{T-t}(\mathbf{x})$, FP becomes $\mathcal{M}v + cv = 0$, where $\mathcal{M}$ is generator of $(Z_t)_{t\in[0,T]}$ and $c : \mathcal{X} \times [0, T] \to \mathbb{R}$.

Model Likelihood

(Generalised) Fokker–Planck PDE

$\partial_t p_t = \hat{\mathcal{K}}^* p_t$

Assumption 1

With $v(\mathbf{x}, t) = p_{T-t}(\mathbf{x})$, FP becomes $\mathcal{M}v + cv = 0$, where $\mathcal{M}$ is generator of $(Z_t)_{t\in[0,T]}$ and $c : \mathcal{X} \times [0, T] \to \mathbb{R}$.

Euclidean Diffusion

Set-up is $\mathcal{K} = \partial_t + \mu \cdot \nabla + \frac{1}{2}\Delta$, and $\mathcal{L} = \partial_t + b \cdot \nabla + \frac{1}{2}\Delta$.
Then, FP PDE is: $\partial_t v = \mu \cdot \nabla v + (\nabla \cdot \mu)v - \frac{1}{2}\Delta v$.
$c = -(\nabla \cdot \mu)$ and $\mathcal{M} = \partial_t - \mu \cdot \nabla + \frac{1}{2}\Delta$.

Model Likelihood

Applying a generalised form of the Feynman–Kac theorem, we can write the model likelihood as

$$p_T(\mathbf{x}) = \mathbb{E}\left[p_0(Z_T) \exp\left\{ \int_0^T c(Z_t, t)\, \mathrm{d}t \right\} \;\middle|\; Z_0 = \mathbf{x} \right]$$

Lower Bound on Model Log Likelihood

Assumption 2

There is $\beta : \mathcal{X} \times [0, T] \to (0, \infty)$ s.t. $\beta^{-1}\mathcal{M}f = \mathcal{L}(\beta^{-1}f) - f\mathcal{L}(\beta^{-1})$.

Recall $\mathcal{K}$ determines $\mathcal{M}$ via $\partial_t p_t = \hat{\mathcal{K}}^* p_t \Leftrightarrow \mathcal{M}v + cv = 0$.

We think of β as parameterising $\mathcal{K}$ via $\mathcal{M}$.

Lower Bound on Model Log Likelihood

Assumption 2

There is $\beta : \mathcal{X} \times [0, T] \to (0, \infty)$ s.t. $\beta^{-1}\mathcal{M}f = \mathcal{L}(\beta^{-1}f) - f\mathcal{L}(\beta^{-1})$.

Recall $\mathcal{K}$ determines $\mathcal{M}$ via $\partial_t p_t = \hat{\mathcal{K}}^* p_t \Leftrightarrow \mathcal{M}v + cv = 0$.

We think of β as parameterising $\mathcal{K}$ via $\mathcal{M}$.

Euclidean Diffusion

Set-up is $\mathcal{K} = \partial_t + \mu \cdot \nabla + \frac{1}{2}\Delta$, and $\mathcal{L} = \partial_t + b \cdot \nabla + \frac{1}{2}\Delta$.
Assumption 2 becomes $\nabla \log \beta = \mu + b$.

Lower Bound on Model Log Likelihood

Starting from

$$\log p_T(\mathbf{x}) = \log \mathbb{E}\left[p_0(Z_T) \exp\left\{ \int_0^T c(Z_t, t)\, \mathrm{d}t \right\} \,\middle|\, Z_0 = \mathbf{x} \right]$$

and applying Jensen's and (generalised) Girsanov,

$$\log p_T(\mathbf{x}) \geq \mathbb{E}_{\mathbb{Q}}\left[\log p_0(Y_T) \middle| Y_0 = \mathbf{x} \right] - \int_0^T \mathbb{E}_{\mathbb{Q}}\left[\frac{\hat{\mathcal{L}}^* \beta}{\beta} + \hat{\mathcal{L}} \log \beta \;\middle|\; Y_0 = \mathbf{x} \right] \mathrm{d}t.$$

Tractable Training Objective

Consider

$$\mathcal{E}^{\infty} := \mathbb{E}_{\mathbb{Q}}\Big[\log p_0(Y_T)\Big|Y_0 = \mathbf{x}\Big] - \int_0^T \mathbb{E}_{\mathbb{Q}}\left[\frac{\hat{\mathcal{L}}^*\beta}{\beta} + \hat{\mathcal{L}}\log\beta \;\middle|\; Y_0 = \mathbf{x}\right]\mathrm{d}t.$$

The first term is constant.

Tractable Training Objective

Consider

$$\mathcal{E}^{\infty} := \mathbb{E}_{\mathbb{Q}}\Big[\log p_0(Y_T)\Big|Y_0 = \mathbf{x}\Big] - \int_0^T \mathbb{E}_{\mathbb{Q}}\left[\frac{\hat{\mathcal{L}}^*\beta}{\beta} + \hat{\mathcal{L}}\log\beta \;\middle|\; Y_0 = \mathbf{x}\right]\mathrm{d}t.$$

The first term is constant. The expectation of the second term is

$$\mathcal{I}_{\mathrm{ISM}}(\beta) = \int_0^T \mathbb{E}_{q_t(\mathbf{x}_t)}\left[\frac{\hat{\mathcal{L}}^*\beta(\mathbf{x}_t,t)}{\beta(\mathbf{x}_t,t)} + \hat{\mathcal{L}}\log\beta(\mathbf{x}_t,t)\right]\mathrm{d}t.$$

This is tractable to minimise!

Tractable Training Objective

We also have the corresponding denoising score matching objective

$$\mathcal{I}_{\mathrm{DSM}}(\beta) = \int_0^T \mathbb{E}_{q_{0,t}} \left[\frac{\mathcal{L}(q_{\cdot|0}/\beta(\cdot,\cdot))(\mathbf{x}_t, t)}{q_{t|0}(\mathbf{x}_t|\mathbf{x}_0)/\beta(\mathbf{x}_t, t)} - \mathcal{L}\log(q_{\cdot|0}/\beta)(\mathbf{x}_t, t) \right] \mathrm{d}t.$$

Tractable Training Objective

We also have the corresponding denoising score matching objective

$$\mathcal{I}_{\mathrm{DSM}}(\beta) = \int_0^T \mathbb{E}_{q_{0,t}} \left[\frac{\mathcal{L}(q_{\cdot|0}/\beta(\cdot,\cdot))(\mathbf{x}_t, t)}{q_{t|0}(\mathbf{x}_t|\mathbf{x}_0)/\beta(\mathbf{x}_t, t)} - \mathcal{L}\log(q_{\cdot|0}/\beta)(\mathbf{x}_t, t) \right] \mathrm{d}t.$$

Euclidean Diffusion

The objective becomes

$$\mathcal{I}_{\mathrm{DSM}}(\beta) = \frac{1}{2}\int_0^T \mathbb{E}_{q_{0,t}(\mathbf{x}_0,\mathbf{x}_t)} \left[||\nabla_{\mathbf{x}} \log q_{t|0}(\mathbf{x}_t|\mathbf{x}_0) - \nabla_{\mathbf{x}} \log \beta(\mathbf{x}_t, t)||^2 \right] \mathrm{d}t.$$

We recover the original diffusion objective.

Other Properties of DMMs

- Can be used for inference; draw $(\mathbf{x}_0, \boldsymbol{\xi}_0) \sim p_{\text{data}}$, noise $\mathbf{x}_0$ according to $\mathcal{L}$, learn generative process conditioned on observation ξ^*, parameterised by $\beta(\mathbf{x}_t, \boldsymbol{\xi}^*, t)$.
- Original discrete-time diffusion model framework of Sohl-Dickstein et al. is natural first order discretisation of DMMs.

Generalised Score Matching

Generalised Score Matching

- $\mathcal{I}_{\mathrm{ISM}}(\beta)$ reduces to implicit score matching objective of Hyvärinen [5] for Euclidean diffusions.

Generalised Score Matching

- $\mathcal{I}_{\mathrm{ISM}}(\beta)$ reduces to implicit score matching objective of Hyvärinen [5] for Euclidean diffusions.
- So, we interpret $\mathcal{I}_{\mathrm{ISM}}(\beta)$ as a generalisation of the score matching objective.

Generalised Score Matching

- $\mathcal{I}_{\mathrm{ISM}}(\beta)$ reduces to implicit score matching objective of Hyvärinen [5] for Euclidean diffusions.
- So, we interpret $\mathcal{I}_{\mathrm{ISM}}(\beta)$ as a generalisation of the score matching objective.
- Given data distribution $q_0(\mathbf{x})$ on $\mathcal{X}$, we learn an approximation $\varphi(\mathbf{x})$ to q_0 by minimising

$$\mathcal{J}_{\mathrm{ESM}}(\varphi) = \mathbb{E}_{q_0(\mathbf{x})}\left[\frac{\mathcal{L}(q_0/\varphi)(\mathbf{x})}{(q_0(\mathbf{x})/\varphi(\mathbf{x}))} - \mathcal{L}\log(q_0/\varphi)(\mathbf{x})\right].$$

Generalised Score Matching

- This is not directly tractable, but is equivalent to

$$\mathcal{J}_{\mathrm{ISM}}(\varphi) = \mathbb{E}_{q_0(\mathbf{x})}\left[\frac{\hat{\mathcal{L}}^*\varphi(\mathbf{x})}{\varphi(\mathbf{x})} + \hat{\mathcal{L}}\log\varphi(\mathbf{x})\right].$$

- This gives a **principled generalisation of score matching to arbitrary state spaces**!
- We define the *score matching operator*

$$\Phi(f) = \frac{\mathcal{L}f}{f} - \mathcal{L}\log f.$$

Generalized Score Matching

Intuitions for score matching on $\mathbb{R}^d$ carry over:

Proposition 1

Feller process Y with generator $\mathcal{L}$, semigroup operators $(Q_t)_{t\geq 0}$ and score matching operator Φ. Then:

1. $\Phi(f) \geq 0$ with equality if f is constant;
2. for probability measures π_1, π_2 on $\mathcal{X}$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(\pi_1 Q_t || \pi_2 Q_t) = -\mathbb{E}_{\pi_1 Q_t}\left[\Phi\left(\frac{\mathrm{d}(\pi_1 Q_t)}{\mathrm{d}(\pi_2 Q_t)}\right)\right].$$

Experimental Performance of DMMs

Discrete Space CTMC: MNIST

We train a DMM to reconstruct images of handwritten digits, conditioned on the border of the image and the value of the digit.

Our state space is $\mathcal{X} = \{0, \ldots, 255\}^{28 \times 28}$ and our noising process is a continuous time Markov chain.

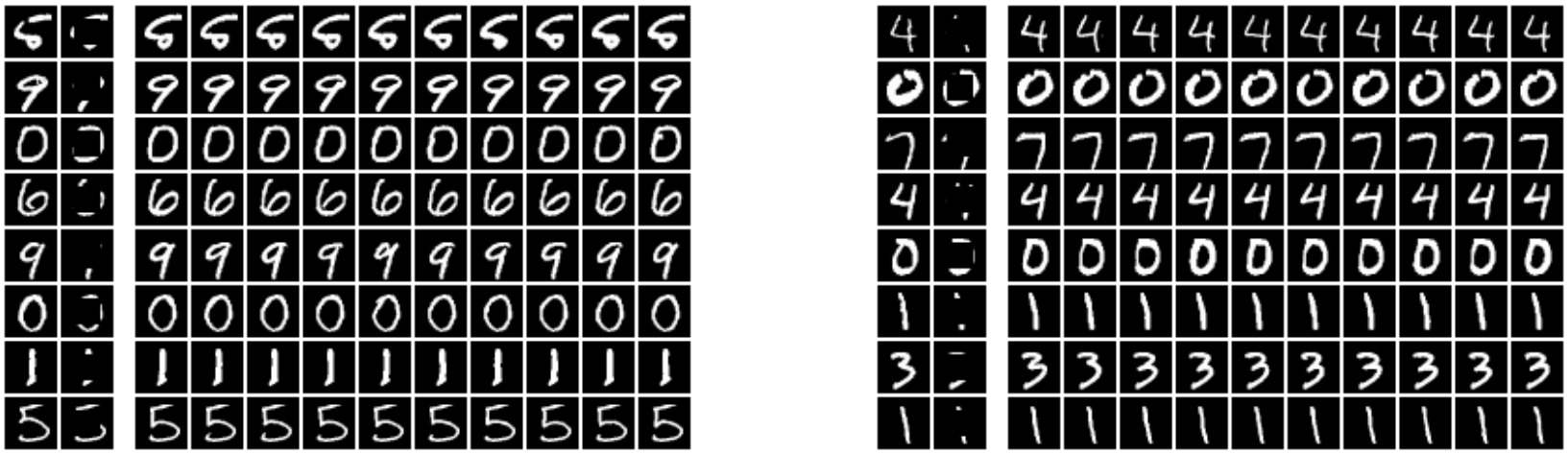


Figure: First column plots the ground truth images. Second column has the centre 14×14 pixels missing.

Brownian Diffusion on $SO(3)$: Pose Estimation

DMM estimates 3D orientation of solids based on 2D views. State space is $\mathcal{X} = SO(3)$, noising process is a Brownian diffusion.

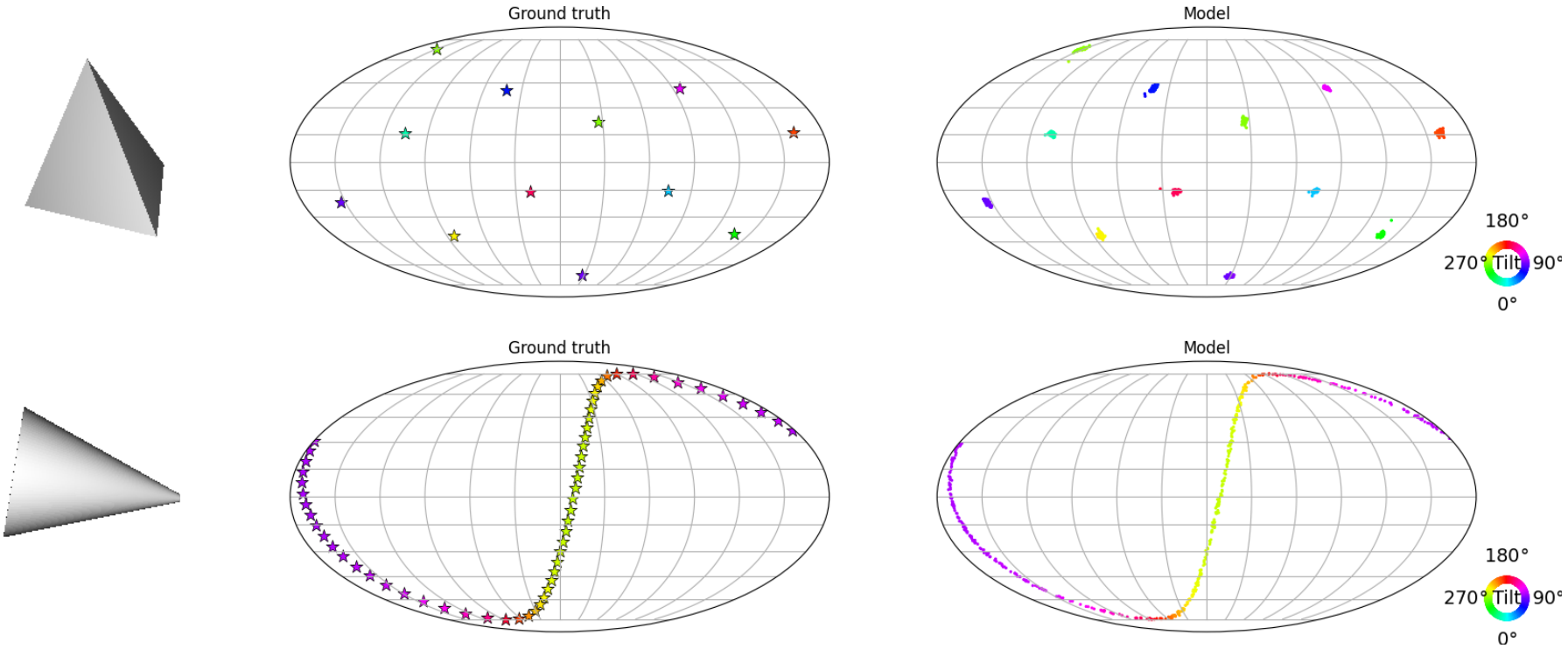


Figure: Ground truth (middle) and DMM estimation (right) of the 3D pose conditioned on 2D views of two shapes (left).

References

- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. *NeurIPS*, 2020.
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv:2204.06125*, 2022.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.
- Chin-Wei Huang, Jae Hyun Lim, and Aaron Courville. A Variational Perspective on Diffusion-Based Generative Models and Score Matching. *NeurIPS*, 2021.
- Aapo Hyvärinen. Estimation of Non-Normalized Statistical Models by Score Matching. *Journal of Machine Learning Research*, 6:695–709, 2005.
- Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A Diffusion Probabilistic Model for Text-to-speech. *ICML*, 2021.
- Brian L Trippe, Jason Yim, Doug Tischer, David Baker, Tamara Broderick, Regina Barzilay, and Tommi Jaakkola. Diffusion Probabilistic Modeling of Protein Backbones in 3D for the Motif-scaffolding Problem. *ICLR*, 2023.