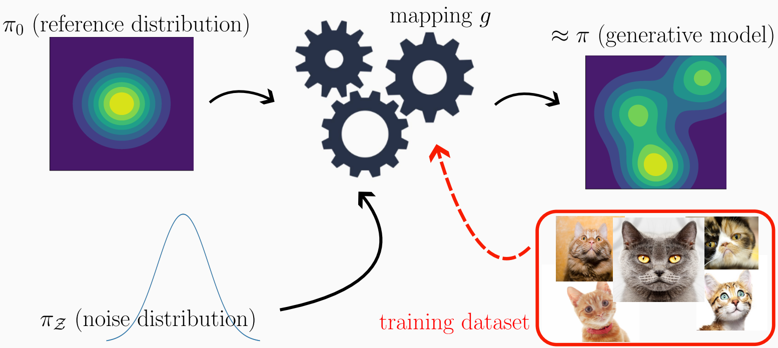
Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling

Valentin De Bortoli

March 16, 2023

What is generative modeling?

- **Generative modeling**: Given a **dataset** of samples from a distribution π how to obtain **new samples** from π?
- **A general approach**:
  - Sample X_0 from π_0 (reference distribution).
  - Sample Z from $\pi_{\mathcal{Z}}$ (noise distribution).
  - Push with $g(X_0, Z) \to$ approximate sample from π.



Why generative modeling?

- Application in **computational biology**: Senior et al. (2020).
  - **Amino-acid sequence** to **3D structure**.
  - Cryo-Electron Microscopy or crystallography = experimental techniques to determine the shape of the protein.
  - Crystallizing a protein is a real challenge Avanzato et al. (2019).
  - Competition to predict structure: **C**ritical **A**ssessment of protein **S**tructure **P**rediction.
- **Conditional generative modeling**.

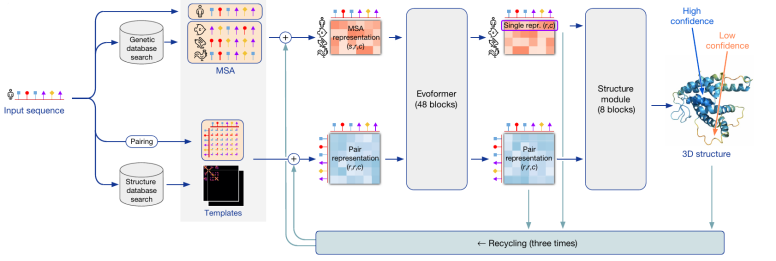
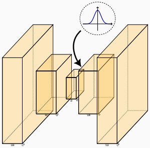


Image extracted from Senior et al. (2020).

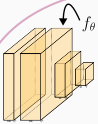
A myriad of models

Variational AutoEncoder



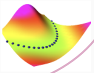
Kingma et al. (2014)
Rezende et al. (2014)
Ranganath et al. (2016)
Vahdat et al. (2021)

Energy-Based Model



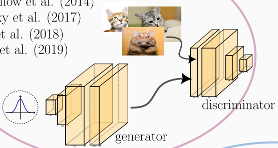
Zhu et al. (1998)
LeCun et al. (2006)
Hinton et al. (2006)
Du et al. (2019)

$$\frac{\exp[-f_\theta(x)]}{\int \exp[-f_\theta(\tilde{x})]\mathrm{d}\tilde{x}}$$



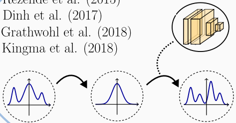
Generative Adversarial Network

Goodfellow et al. (2014)
Arjovsky et al. (2017)
Brock et al. (2018)
Karras et al. (2019)



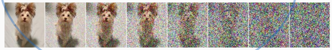
Normalizing Flow

Rezende et al. (2015)
Dinh et al. (2017)
Grathwohl et al. (2018)
Kingma et al. (2018)

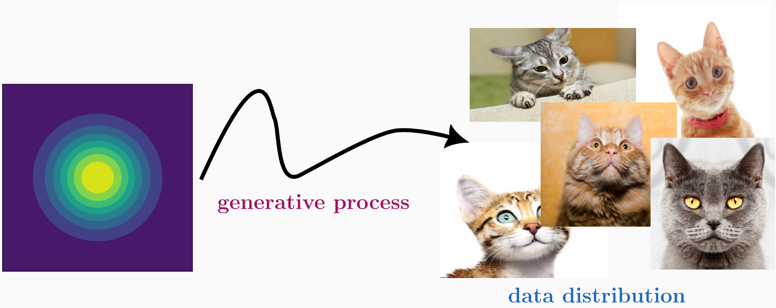


Denoising Diffusion Model

Song et al. (2019)
Ho et al. (2020)
Vahdat et al. (2021)



Some challenges in generative modeling



Theoretical understanding

- Convergence of generative models?

Properties of the data

- Riemannian data.
- Inverse problems.

Properties of the process

- Optimal transport.
- Stochastic control.

Focus on denoising diffusion models.

Generative Modeling: the rise of diffusion models

Time-reversal of diffusions

- **Forward decomposition**: $p(x_{0:N}) = p_0(x_0) \prod_{k=0}^{N-1} p_{k+1|k}(x_{k+1}|x_k)$.

- **Backward decomposition**: $p(x_{0:N}) = p_N(x_N) \prod_{k=0}^{N-1} p_{k|k+1}(x_k|x_{k+1})$.

Video extracted from Song and Ermon (2019).

Approximate time reversal

¿How to approximate the backward decomposition?

- **Backward decomposition**: $p(x_{0:N}) = p_N(x_N) \prod_{k=0}^{N-1} p_{k|k+1}(x_k|x_{k+1})$.
  - How to compute $p_{k|k+1}(x_k|x_{k+1}) = p_{k+1|k}(x_{k+1}|x_k)p_k(x_k)/p_{k+1}(x_{k+1})$?
  - In practice $p_{k+1|k} = \mathrm{N}(x_k - \gamma x_k, \sqrt{2\gamma}\,\mathrm{Id})$ is **Gaussian**.
  - (**Discretization** of $\mathrm{d}\mathbf{X}_t = -\mathbf{X}_t \mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t$ (**Ornstein-Ulhenbeck**))
  - $p_{k|k+1}$ is approximately Gaussian

$$p_{k|k+1} = \mathrm{N}(x_{k+1} + \gamma x_{k+1} + 2\gamma \nabla \log p_{k+1}(x_{k+1}), \sqrt{2\gamma}\,\mathrm{Id}).$$

¿How to compute the score term?

- **Score matching** techniques: Vincent (2011); Hyvärinen (2005)

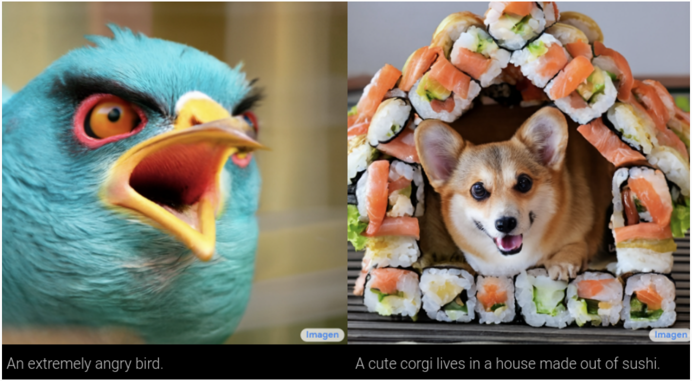
$$\nabla \log p_{k+1}(x_{k+1}) = \mathbb{E}_{p_{0|k+1}}[\nabla \log p_{k+1|0}(x_{k+1}|X_0)].$$

  - **Loss function**: $\ell(\mathbf{s}_{k+1}) = \mathbb{E}[\|\mathbf{s}_{k+1}(X_{k+1}) - \nabla \log p_{k+1|0}(X_{k+1}|X_0)\|^2]$.
  - Algorithm: replace $\nabla \log p_{k+1}$ by $\mathbf{s}_{k+1}$.

Unconditional CelebA synthesis

An application: text-to-image

- From prompt to images: Imagen, DALL-E 2, Stable Diffusion, Midjourney.



- **CLIP** (**C**ontrastive **L**anguage–**I**mage **P**re-training) guidance.

Convergence of diffusion models ($\hat{\pi}$)

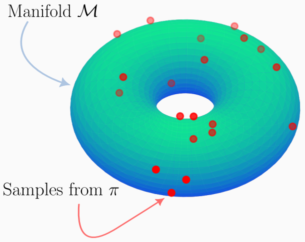
Under dissipativity conditions (D.B et al., 2021[1])

- $\|\mathbf{s}_t(x) - \nabla \log p_t(x)\| \leq \mathtt{M}$.
- π admits a density p and $\langle \nabla \log p(x), x \rangle \leq -\mathtt{m}\|x\|^2 + \mathtt{c}$.

■ Then, there exists $A \geq 0$ such that

forward convergence — discretization — score approximation

$$\|\pi - \hat{\pi}\|_{\mathrm{TV}} \leq A(\exp[-T] + \exp[T](\gamma^{1/2} + \mathtt{M}))$$



Under the manifold hypothesis (D.B., 2022[2])

- π is supported on a compact manifold $\mathcal{M}$.

■ Then there exists $A \geq 0$ such that

$$\mathbf{W}_1(\pi, \hat{\pi}) \leq A(\exp[-T] + \gamma^{1/2} + \mathtt{M}).$$

[1] **D.B.**, Thornton, Heng, Doucet – Diffusion Schrödinger Bridge – NeurIPS 2021
[2] **D.B.** – Convergence of diffusion models under manifold hypotheses – TMLR 2022

Convergence of diffusion models

A more precise statement

Convergence result under the manifold hypothesis (D.B., 2022[3])

Under the manifold hypothesis and controls on the score approximation, there exists $\mathrm{D}_0 \geq 0$ such that

$$\boxed{\mathbf{W}_1(\hat{\pi}, \pi) \leq \mathrm{D}_0(\exp[\kappa/\varepsilon](\mathtt{M} + \gamma^{1/2})/\varepsilon^2 + \exp[\kappa/\varepsilon]\exp[-T/\bar{\beta}] + \varepsilon^{1/2}) \ ,}$$

with $\kappa = \mathrm{diam}(\mathcal{M})^2(1 + \bar{\beta})/2$ and D_0 an explicit constant.

- We control three terms:
  - **Discretization term**: $\mathtt{M}$, network error ; γ, discretization stepsize.
  - **Convergence term**: T, forward time.
  - **Non-degeneracy term**: ε, stopping time in the backward.
- First, we discuss the assumptions:
  - Manifold hypothesis (assumption on π).
  - Score approximation (assumption on $\mathbf{s}_\theta$).

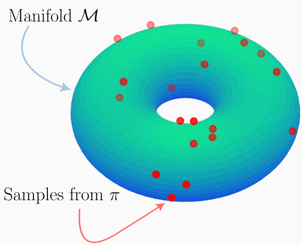
[2] **D.B.** – Convergence of diffusion models under manifold hypotheses – TMLR 2022

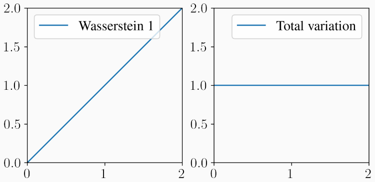
Distances and the manifold hypothesis

- Problem with **total variation distance**:
  - μ, ν with disjoint supports, $\|\mu - \nu\|_{\mathrm{TV}} = 1$.
  - No notion of **sample proximity** ("vertical" distance).
- The **manifold hypothesis**:
  - Data distribution is supported on a **low-dimensional** compact space $\mathcal{M} \subset \mathbb{R}^d$.
  - However, generative model has distribution on $\mathbb{R}^d$.
  - Under the manifold hypothesis

$$\boxed{\|\pi - \hat{\pi}\|_{\mathrm{TV}} = 1\ .}$$

- Let's turn to **Wasserstein distances** ("horizontal distance").

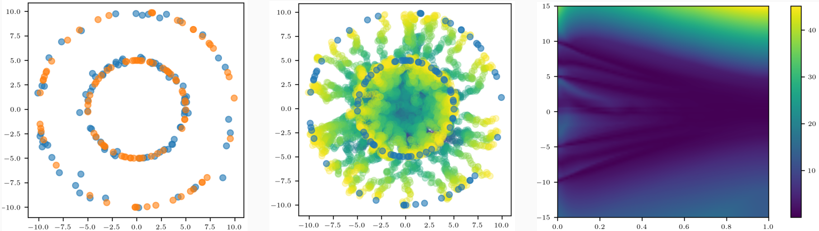




Assumption on the score

Uniform control on the score

There exists $\mathtt{M} \geq 0$ such that $\|s_\theta(t, x_t) - \nabla \log p_t(x_t)\| \leq \mathtt{M}(1 + \|x_t\|)/\sigma_t^2$



- **Uniform** assumption but allows for **explosive** behavior.
- Behaviour observed in practice.
- More realistic assumptions (L^2 error) in Chen et al. (2022); Lee et al. (2022).

Other assumptions and special ingredient

- The diffusion is usually given with a **speed**

$$\boxed{\mathrm{d}\mathbf{X}_t = -\beta_t \mathbf{X}_t \mathrm{d}t + \sqrt{2\beta_t}\mathrm{d}\mathbf{B}_t \ .}$$

- $\beta_0 \ll \beta_T$ in practice and **linear schedule**.

Control of the speed

$t \mapsto \beta_t$ is continuous, non-decreasing and there exists $\bar{\beta} > 0$ such that for any $t \in [0, T]$, $1/\bar{\beta} \leq \beta_t \leq \bar{\beta}$.

- Control of the stepsize.

Control of the stepsize

For any $k \in \{0, \dots, K-1\}$, we have $\gamma_k \sup_{v \in [T-t_{k+1}, T-t_k]} \beta_v / \sigma_v^2 \leq \gamma \leq 1/2$.

- Satisfied if γ_k **small enough**.
- To avoid **degeneracy**, we *do not* consider the last step (as in Song et al. (2020)).

The central decomposition

- The central decomposition

$$\mathbf{W}_1(\pi_\infty \mathrm{R}_K, \pi) \leq \mathbf{W}_1(\pi_\infty \mathrm{R}_K, \pi_\infty \mathrm{Q}_{t_K}) + \mathbf{W}_1(\pi_\infty \mathrm{Q}_{t_K}, \pi \mathrm{P}_{T-t_K}) + \mathbf{W}_1(\pi \mathrm{P}_{T-t_K}, \pi)\ .$$

  where

  - $(\mathrm{P}_t)_{t \geq 0}$ is the **forward** Ornstein-Ulhenbeck semi-group,
  - $(\mathrm{Q}_t)_{t \geq 0}$ is the **backward** Ornstein-Ulhenbeck semi-group,
  - $(\mathrm{R}_k)_{k \in \{0,\ldots,K-1\}}$ is the iterated kernel associated with the backward Markov chain.

- Decomposition of the error:
  - **Discretization term**: $\mathbf{W}_1(\pi_\infty \mathrm{R}_K, \pi_\infty \mathrm{Q}_{t_K})$.
  - **Convergence term**: $\mathbf{W}_1(\pi_\infty \mathrm{Q}_{t_K}, \pi \mathrm{P}_{T-t_K})$.
  - **Non-degeneracy term**: $\mathbf{W}_1(\pi \mathrm{P}_{T-t_K}, \pi)$.

Controlling the discretization

- Problem with the **Wasserstein distance**:
  - Do not satisfy $\mathbf{W}_1(\mu \mathrm{Q}, \nu \mathrm{Q}) \leq \mathbf{W}_1(\mu, \nu)$.
  - We have to **control the backward**.
- Control of the backward process:
  - Use of the interpolation formula del Moral and Singh (2019)

$$\mathrm{d}\mathbf{Y}^x_{s,t} = \beta_{T-t}\{\mathbf{Y}^x_{s,t} + 2\nabla \log q_{T-t}(\mathbf{Y}^x_{s,t})\}\mathrm{d}t + \sqrt{2\beta_{T-t}}\mathrm{d}\mathbf{B}_t \,, \qquad \mathbf{Y}^x_{s,s} = x \,.$$

$$\mathrm{d}\bar{\mathbf{Y}}^x_{s,t} = \beta_{T-t}\{\bar{\mathbf{Y}}^x_{s,t} + 2\mathbf{s}_\theta(T - t_k, \bar{\mathbf{Y}}^x_{s,t_k})\}\mathrm{d}t + \sqrt{2\beta_{T-t}}\mathrm{d}\mathbf{B}_t \,, \qquad \bar{\mathbf{Y}}^x_{s,s} = x \,.$$

$$\boxed{\mathbf{Y}^x_{s,t} - \bar{\mathbf{Y}}^x_{s,t} = \int_s^t \nabla \mathbf{Y}_{u,t}(\bar{\mathbf{Y}}_{s,u})^\top \Delta b_u((\bar{\mathbf{Y}}_{s,v})_{v \in [s,u]})\mathrm{d}u \,,}$$

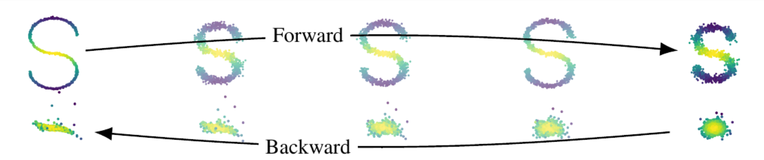
  - Uniform control of the **tangent process** $(\nabla \mathbf{Y}_{u,t})_{u,t \in [0,T]}$.
  - **Explosion** of the score near time 0 (observed in practice!).
- Solution? **Stop** the process before time 0 (at time ε, done in practice).

$$\boxed{\mathbf{W}_1(\hat{\pi}, \pi) \leq D(\exp[\kappa/\varepsilon](\mathrm{M} + \delta^{1/2})/\varepsilon^2 + \exp[\kappa/\varepsilon]\exp[-T/\bar{\beta}] + \varepsilon^{1/2}) \,.}$$

Schrödinger Bridges: a new generative modeling framework

Shorter generative processes?

- **Not enough stepsizes** leads to poor approximation (the Ornstein-Ulhenbeck process does not mix fast enough).



- Illustration of failure: N is too small so p_N is very different from p_{prior}. This harms the quality of the reconstruction for the time-reversal.

- **Trade-off**:
  - Large $N \to$ improvement in **quality** (fidelity).
  - Large $N \to$ **model is slow** at sampling time.

Challenge: how to “shorten” the diffusion process?

The trilemma of generative modeling

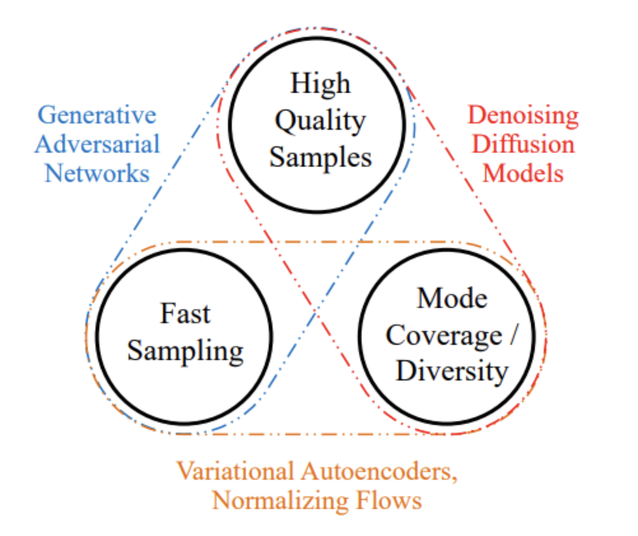


Image extracted from Xiao et al. (2021).

Revisiting Generative Modeling using Schrödinger Bridges

- The **Schrödinger Bridge (SB) problem** is a classical problem appearing in applied mathematics, optimal transport and probability.
  - Consider a **reference density** $p(x_{0:N})$, find $\pi^\star(x_{0:N})$ such that

$\pi^\star$ distribution on $(\mathbb{R}^d)^{N+1}$

$$\pi^\star = \arg\min\{\mathrm{KL}(\pi|p) : \pi_0 = p_{\text{data}},\ \pi_N = p_{\text{prior}}\}.$$

  - **Goal:** If $\pi^\star$ is available: $X_N \sim p_{\text{prior}}$ and $X_k \sim \pi^\star_{k|k+1}(\cdot|X_{k+1})$.

- **Static formulation**: $\pi^\star(x_{0:N}) = \pi^{s,\star}(x_0, x_N)p_{|0,N}(x_{1:N-1}|x_0, x_N)$ where
  - Variational form:

$\pi^{s,\star}$ distribution on $(\mathbb{R}^d)^2$

$$\pi^{s,\star} = \arg\min\{\mathrm{KL}(\pi^s|p_{0,N}) : \pi^s_0 = p_{\text{data}},\ \pi^s_N = p_{\text{prior}}\}.$$

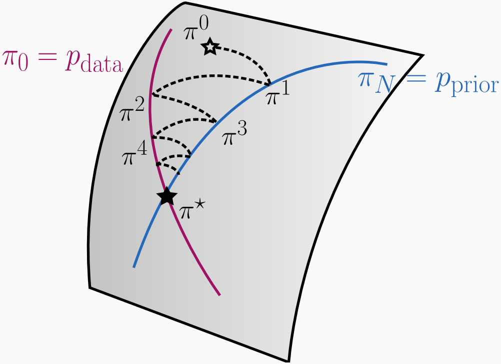
  - In its static form the Schrödinger Bridge is a special case of **entropic optimal transport**, see Mikami (2004).

The Iterative Proportional Fitting algorithm

- The SB problem can be solved using **Iterative Proportional Fitting (IPF)** Sinkhorn and Knopp (1967); Fortet (1940), i.e. set $\pi^0 = p$ and for $n \in \mathbb{N}$

$$\pi^{2n+1} = \arg\min\{\mathrm{KL}(\pi|\pi^{2n}), \ \ \pi_N = p_{\mathrm{prior}}\},$$
$$\pi^{2n+2} = \arg\min\{\mathrm{KL}(\pi|\pi^{2n+1}), \ \ \pi_0 = p_{\mathrm{data}}\}.$$

- This is akin to **alternative projection** in a Euclidean setting.
- $\lim_{n\to+\infty} \pi^n = \pi^\star$ under regularity conditions.



Solving the Schrödinger Bridge

- **Explicit solution** of the first IPF step

$$\mathrm{KL}(\pi|\pi^0) = \mathrm{KL}(\pi_N|p_N) + \mathbb{E}_{\pi_N}[\mathrm{KL}(\pi_{|N}|p_{|N})].$$

Therefore,

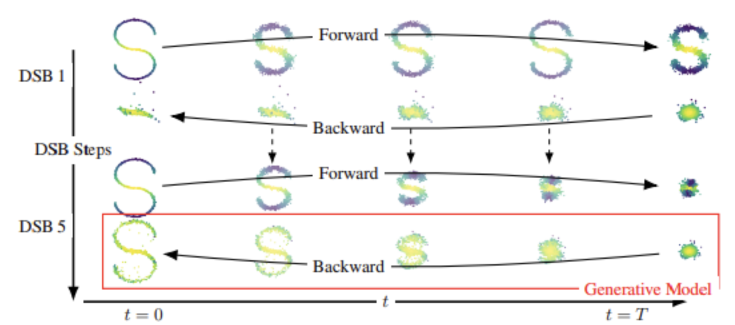
$$\pi^1(x_{0:N}) = p_{\text{prior}}(x_N)p(x_{0:N-1}|x_N)$$

$$\boxed{\pi^1(x_{0:N}) = p_{\text{prior}}(x_N)\prod_{k=0}^{N-1} p_{k|k+1}(x_k|x_{k+1}).}$$

- **Take-home message:** Approximation to first iteration of IPF corresponds to current **denoising diffusion models**.
- The IPF is a **refinement** on denoising diffusion models.

Diffusion Schrödinger Bridge

- **Diffusion Schrödinger Bridge**[4]:
  - Use **diffusion models** to solve IPF at each step.
  - Alternate between updating the **forward** and **backward dynamics**.
  - (One network parameterizing the forward, one parameterizing the backward).



[4] **D.B.**, Thornton, Heng, Doucet – Diffusion Schrödinger Bridge – NeurIPS 2021

2D illustration

Conclusion

Conclusion

- Fruitful interaction between **stochastic processes** and **generative modeling**.

- Extension to other data/process constraints built on **stochastic processes**.

- Promising developments of **control** and **optimal transport** techniques for generative models (and vice-versa).

"Thank you" generated with the text-to-prompt model Stable diffusion.

References

Victoria A Avanzato, Kasopefoluwa Y Oguntuyo, Marina Escalera-Zamudio, Bernardo Gutierrez, Michael Golden, Sergei L Kosakovsky Pond, Rhys Pryce, Thomas S Walter, Jeffrey Seow, Katie J Doores, et al. A structural basis for antibody-mediated neutralization of nipah virus reveals a site of vulnerability at the fusion glycoprotein apex. *Proceedings of the National Academy of Sciences*, 116(50):25057–25067, 2019.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2022.

Pierre del Moral and Sumeetpal Sidhu Singh. Backward it {\ˆ o}-ventzell and stochastic interpolation formulae. *arXiv preprint arXiv:1906.09145*, 2019.

Robert Fortet. Résolution d'un système d'équations de M. Schrödinger. *Journal de Mathématiques Pures et Appliqués*, 1:83–105, 1940.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. *arXiv preprint arXiv:2206.06227*, 2022.

Toshio Mikami. Monge's problem with a quadratic cost by the zero-noise limit of h-path processes. *Probability theory and related fields*, 129(2):245–260, 2004.

Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, 2021.

Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. *arXiv preprint arXiv:2112.07804*, 2021.

Approximating Backward Transitions

- We restrict ourselves to discretized **Ornstein-Ulhenbeck** processes

$$p_{k+1|k}(x_{k+1}|x_k) = \mathcal{N}(x_{k+1}; x_k - \gamma x_k, \sqrt{\gamma}\,\mathrm{Id}),$$

($\gamma > 0$ is close to 0)

- Using a Taylor expansion we get

$$p_{k|k+1}(x_k|x_{k+1}) = p_{k+1|k}(x_{k+1}|x_k)\exp[\log p_k(x_k) - \log p_{k+1}(x_{k+1})]$$
$$\approx \mathcal{N}(x_k; x_{k+1} + \gamma x_{k+1} + 2\gamma \underbrace{\nabla \log p_{k+1}(x_{k+1})}_{\text{Stein score}}, \sqrt{2\gamma}\,\mathrm{Id}).$$

- The **Stein score** is not available but using that $p_{k+1}(x_{k+1}) = \int p_0(x_0) p_{k+1|0}(x_{k+1}|x_0)\mathrm{d}x_0$, we get that

$$\boxed{\nabla \log p_{k+1}(x_{k+1}) = \mathbb{E}_{p_{0|k+1}}[\nabla_{x_{k+1}} \log p_{k+1|0}(x_{k+1}|X_0)].}$$

Estimating the Scores using Score Matching

- **Conditional expectation** → **Regression problem**

$$s_{k+1} = \arg\min_s \ \mathbb{E}_{p_{0,k+1}}[||s(X_{k+1}) - \nabla_{x_{k+1}} \log p_{k+1|0}(X_{k+1}|X_0)||^2].$$

- In practice, we restrict ourselves to **neural networks** and estimate all scores simultaneously i.e. $s_{\theta^\star}(k, x_k) \approx \nabla \log p_k(x_k)$ where

$$\boxed{\theta^\star \approx \arg\min_\theta \sum_{k=1}^N \mathbb{E}_{p_{0,k}}[||s_\theta(k, X_k) - \nabla_{x_k} \log p_{k|0}(X_k|X_0)||^2],}$$

- If $\log p_{k+1|0}(x_{k+1}|x_0)$ is not available, then use

$$\nabla \log p_{k+1}(x_{k+1}) = \mathbb{E}_{p_{k|k+1}}[\nabla_{x_{k+1}} \log p_{k+1|k}(x_{k+1}|X_k)]$$

- Can also be derived from a **continuous-time** perspective (time-reversal of diffusion, Feynman-Kac formula) and can be seen as ELBO (Huang et al., 2021).
- Yet another approach goes fully variational (Ho et al., 2020).

Sketch of the proof

- The central decomposition

$$\begin{aligned}||\mathcal{L}(X_0) - p_{\text{data}}||_{\text{TV}} &= ||p_{\text{prior}}\hat{\mathrm{R}}_N - p_{\text{data}}||_{\text{TV}} \\ &= ||p_{\text{prior}}\hat{\mathrm{R}}_N - p_T \mathrm{Q}_T||_{\text{TV}} \\ &\leq ||p_{\text{prior}}\hat{\mathrm{R}}_N - p_{\text{prior}}\mathrm{Q}_T||_{\text{TV}} + ||p_T\mathrm{Q}_T - p_{\text{prior}}\mathrm{Q}_T||_{\text{TV}} \\ &\leq ||p_{\text{prior}}\hat{\mathrm{R}}_N - p_{\text{prior}}\mathrm{Q}_T||_{\text{TV}} + ||p_{\text{data}}\mathrm{P}_T - p_{\text{prior}}||_{\text{TV}},\end{aligned}$$

  where
  - $(\mathrm{P}_t)_{t\geq 0}$ is the **forward** Ornstein-Ulhenbeck semi-group,
  - $(\mathrm{Q}_t)_{t\geq 0}$ is the **backward** Ornstein-Ulhenbeck semi-group,
  - $(\hat{\mathrm{R}}_n)_{n\in\{1,\ldots,N\}}$ is the iterated kernel associated with the backward Markov chain.

- $||p_{\text{prior}}\hat{\mathrm{R}}_N - p_{\text{prior}}\mathrm{Q}_T||_{\text{TV}}$: **approximation error** $\to$ Girsanov theorem.

- $||p_{\text{data}}\mathrm{P}_T - p_{\text{prior}}||_{\text{TV}}$: **geometric ergodicity** of Ornstein-Ulhenbeck.

Reverse process on a compact manifold

- The **Brownian motion** is defined as a process $(\mathbf{B}_t^{\mathcal{M}})_{t \geq 0}$ such that for any $f \in \mathrm{C}^\infty(\mathcal{M})$, $(\mathbf{M}_t^f)_{t \geq 0}$ is a martingale where for any $t \geq 0$

$$\boxed{\mathbf{M}_t^f = f(\mathbf{B}_t^{\mathcal{M}}) - f(\mathbf{B}_0^{\mathcal{M}}) - \int_0^t (1/2) \Delta_{\mathcal{M}}(f)(\mathbf{B}_s^{\mathcal{M}}) \mathrm{d}s.}$$

- The **reverse process** is given by $(\mathbf{Y}_t)_{t \in [0,T]}$ such that for any $f \in \mathrm{C}^\infty(\mathcal{M})$, $(\mathbf{M}_t^f)_{t \geq 0}$ is a martingale where for any $t \in [0, T]$

$$\boxed{\mathbf{M}_t^f = f(\mathbf{Y}_t) - f(\mathbf{Y}_0) - \int_0^t \{ \langle \nabla \log p_t(\mathbf{X}_s), \nabla f(\mathbf{Y}_s) \rangle_{\mathcal{M}} + (1/2) \Delta_{\mathcal{M}}(f)(\mathbf{Y}_s) \} \mathrm{d}s.}$$

- This is an extension of **reversal** results (Haussmann et al., 1986) (Conforti et al., 2021).
- **Take-home message:** The formula is the same except that **gradients, scalar product and Laplacian** are considered w.r.t. the underlying metric.

Sampling on a manifold

- How to sample from the process $(\mathbf{Y}_t)_{t\in[0,T]}$ (approximately)?
- Equivalent of the **Euler-Maruyama** discretization is the **Geodesic Random Walk** (GRW)

Definition of GRW

Let X_0^γ be a $\mathcal{M}$-valued random variable. For any $\gamma > 0$, we define $(X_n^\gamma)_{n\in\mathbb{N}}$ such that for any $n \in \mathbb{N}$,

$$\boxed{X_{n+1}^\gamma = \exp_{X_n^\gamma}\left(\gamma\{b(X_n^\gamma) + (1/\sqrt{\gamma})(V_{n+1} - b(X_n^\gamma))\}\right).}$$

where $(V_n)_{n\in\mathbb{N}}$ is a sequence of $\mathcal{M}$-valued random variables such that for any $n \in \mathbb{N}$, V_{n+1} has distribution $\nu_{X_n^\gamma}$ conditionally to X_n^γ (mean $b(X_n^\gamma)$, covariance $\Sigma(X_n^\gamma)$).

- **Weakly converges** towards the diffusion $\mathrm{d}\mathbf{X}_t = b(\mathbf{X}_t)\mathrm{d}t + \Sigma(\mathbf{X}_t)\mathrm{d}\mathbf{B}_t^{\mathcal{M}}$ for small stepsizes γ.
- Hard to obtain **quantitative results** (coupling techniques in Riemannian setting).

Perspectives & Challenges

Plan

Some challenges:

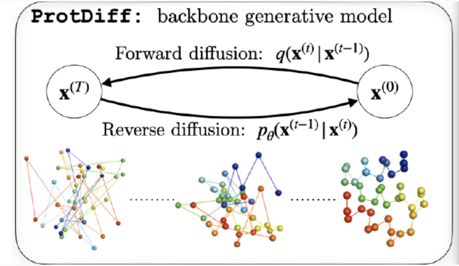
- **Scaling up** Diffusion Schrodinger Bridge and protein applications.
- Particle evolution and **probabilistic splines**.
- **Theoretical understanding** of diffusion models and other projects.

Scaling up and protein applications

- To be competitive: access to large **GPU infrastructure**.

ImageNet 512×512							
BigGAN-deep [5]			256-512	8.43	8.13	**0.88**	0.29
ADM-G (4360K), ADM-U (1050K)	1878	36	1914	**3.85**	**5.86**	0.84	**0.53**
ADM-G (500K), ADM-U (100K)	189	9*	**198**	7.59	6.84	0.84	**0.53**

- More than **200** V100 days to train one SoTA diffusion model on ImageNet 512×512.
- Importance of the scaling for:
  - **Image processing** (realistic outputs, interaction with language models...)
  - **Protein Modeling** (long proteins...) (image from Trippe et al. (2022))



Particle evolution and spline

- For **population evolution**, one Schrödinger bridge is not enough.
- **Multiple snapshots**, can we consider multiple Schrödinger bridges?
- How can we impose some regularity in the **probabilistic structure**?
  - **Spline** in probabilistic spaces (Chen et al. (2018))
  - Efficient combination with Diffusion Schrödinger Bridges.

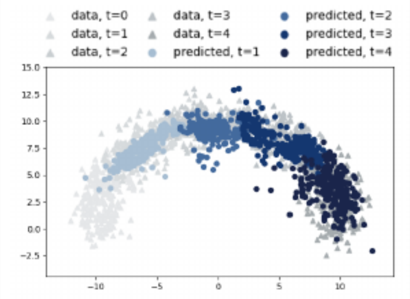


Image extracted from Bunne et al. (2022)

Theoretical understanding of diffusion models & other projects

- A lot of **open questions**:
  - Role of the **manifold hypothesis**.
  - Role of the **Empirical measure**.
  - And what about **multimodal** behavior?

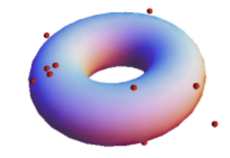
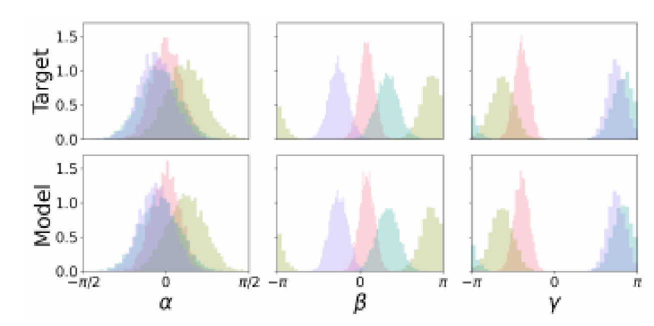


Image extracted from Fefferman et al. (2015)

- Other projects
  - **VAE** as **entropic regularization**
  - Interpretation of **Transformers** with **category theory** tools.

Some results on $SO_3(\mathbb{R})$

- An illustration: targeting **multimodal distributions** on $SO_3(\mathbb{R})$.



Method	$M = 16$		$M = 32$	
	log-likelihood	NFE	log-likelihood	NFE
Moser Flow	$0.85_{\pm 0.03}$	$2.3_{\pm 0.5}$	$0.17_{\pm 0.03}$	$2.3_{\pm 0.9}$
Exp-wrapped SGM	$\mathbf{0.87}_{\pm 0.04}$	$0.5_{\pm 0.1}$	$0.16_{\pm 0.03}$	$0.5_{\pm 0.0}$
RSGM	$\mathbf{0.89}_{\pm 0.03}$	$\mathbf{0.1}_{\pm 0.0}$	$\mathbf{0.20}_{\pm 0.03}$	$\mathbf{0.1}_{\pm 0.0}$

Motivation

- Many datasets do *not* lie on a **Euclidean space**.
- We need to include a **geometric prior**:
  - **Protein modeling** (Boomsma et al., 2008; Hamelryck et al., 2006; Mardia et al., 2008; Shapovalov and Dunbrack Jr, 2011; Mardia et al., 2007).
  - **Geological sciences** (Karpatne et al., 2018; Peel et al., 2001).
  - **Robotics** (Feiten et al., 2013; Senanayake and Ramos, 2018).

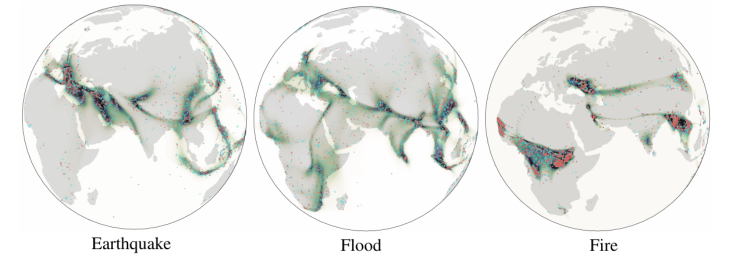


Image extracted from Mathieu et al., 2020.

Noising process on a compact manifold

- To define a **score-based generative modeling** we need to define a **noising process**
  - In **Euclidean spaces** we choose a **Ornstein-Ulhenbeck** process.
  - In **Riemannian manifold** we choose a **Brownian motion**.
- In the **Euclidean** setting the **Ornstein-Ulhenbeck** process converges towards a unit Gaussian.
- In the *compact* **Riemannian manifold** setting the **Brownian motion** converges towards the uniform distribution.

Geometric ergodicity (Urakawa, 2006, Proposition 2.6)

For any $t > 0$, P_t admits a density $p_{t|0}$ w.r.t. p_{ref} and $p_{\mathrm{ref}} \mathrm{P}_t = p_{\mathrm{ref}}$, *i.e.* p_{ref} is an invariant measure for $(\mathrm{P}_t)_{t \geq 0}$. In addition, if there exists $C, \alpha \geq 0$ such that $p_{t|0}(x|x) \leq C t^{-\alpha/2}$ for any $t \in (0, 1]$ and any $x \in \mathcal{M}$ then for any $p_0 \in \mathcal{P}(\mathcal{M})$ and for any $t \geq 1/2$ we have

$$\| p_0 \mathrm{P}_t - p_{\mathrm{ref}} \|_{\mathrm{TV}} \leq C^{1/2} \mathrm{e}^{\lambda_1/2} \mathrm{e}^{-\lambda_1 t},$$

where λ_1 is the first non-negative eigenvalue of $-\Delta_{\mathcal{M}}$ in $\mathrm{L}^2(p_{\mathrm{ref}})$.

Reverse process on a compact manifold

- The **Brownian motion** is defined as a process $(\mathbf{B}_t^{\mathcal{M}})_{t \geq 0}$ such that for any $f \in \mathrm{C}^\infty(\mathcal{M})$, $(\mathbf{M}_t^f)_{t \geq 0}$ is a martingale where for any $t \geq 0$

$$\mathbf{M}_t^f = f(\mathbf{B}_t^{\mathcal{M}}) - f(\mathbf{B}_0^{\mathcal{M}}) - \int_0^t (1/2)\Delta_{\mathcal{M}}(f)(\mathbf{B}_s^{\mathcal{M}})\mathrm{d}s.$$

- The **reverse process** is given by $(\mathbf{Y}_t)_{t \in [0,T]}$ such that for any $f \in \mathrm{C}^\infty(\mathcal{M})$, $(\mathbf{M}_t^f)_{t \geq 0}$ is a martingale where for any $t \in [0, T]$

$$\boxed{\mathbf{M}_t^f = f(\mathbf{Y}_t) - f(\mathbf{Y}_0) - \int_0^t \{\langle \nabla_{\mathcal{M}} \log p_t(\mathbf{X}_s), \nabla_{\mathcal{M}} f(\mathbf{Y}_s)\rangle_{\mathcal{M}} + (1/2)\Delta_{\mathcal{M}}(f)(\mathbf{Y}_s)\}\mathrm{d}s.}$$

- This is an extension of **reversal** results (Haussmann et al., 1986) (Conforti et al., 2021).
- The formula is the same except that **gradients, scalar product and Laplacian** are considered w.r.t. the underlying metric.

Sampling on a manifold

- How to sample from the process $(bfY_t)_{t\in[0,T]}$ (approximately)?
- Equivalent of the **Euler-Maruyama** discretization is the **Geodesic Random Walk** (GRW)

Definition of GRW

Let X_0^γ be a $\mathcal{M}$-valued random variable. For any $\gamma > 0$, we define $(X_n^\gamma)_{n\in\mathbb{N}}$ such that for any $n \in \mathbb{N}$, $X_{n+1}^\gamma = \exp_{X_n^\gamma}\left(\gamma\{b(X_n^\gamma) + (1/\sqrt{\gamma})(V_{n+1} - b(X_n^\gamma))\}\right)$, where $(V_n)_{n\in\mathbb{N}}$ is a sequence of $\mathcal{M}$-valued random variables such that for any $n \in \mathbb{N}$, V_{n+1} has distribution $\nu_{X_n^\gamma}$ conditionally to X_n^γ (mean $b(X_n^\gamma)$, covariance $\Sigma(X_n^\gamma)$).

Convergence of GRW (Jorgensen, 1975, Theorem 2.1)

Under mild conditions on $\mathcal{M}$, for any $t \geq 0$, $f \in \mathrm{C}(\mathcal{M})$ we have that $\lim_{\gamma\to 0}\left|\mathbb{E}\left[f(X_{\lceil t/\gamma\rceil}^\gamma)\right] - \mathrm{P}_t[f]\right| = 0$, where $(\mathrm{P}_t)_{t\geq 0}$ is the semi-group associated with the infinitesimal generator $\mathscr{A} : \mathrm{C}^\infty(\mathcal{M}) \to \mathrm{C}^\infty(\mathcal{M})$ given for any $f \in \mathrm{C}^\infty(\mathcal{M})$ by $\mathscr{A}(f) = \langle b, \nabla f\rangle_{\mathcal{M}} + \frac{1}{2}\langle \Sigma, \nabla^2 f\rangle_{\mathcal{M}}$.

- Hard to obtain **quantitative results** (coupling techniques fail).

Loss function

- We need to estimate $\nabla \log p_t$.
- Same as **Euclidean** case, $\nabla \log p_t(x_t) = \mathbb{E}[\nabla \log p_{t|0}(\mathbf{X}_t|\mathbf{X}_0)|\mathbf{X}_t = x_t]$.
- Extra difficulty, $\nabla \log p_{t|0}$ is *not* available in **close form**.
- Two possibilities to circumvent this issue:
  - Use the **divergence theorem**

$$\boxed{\nabla \log p_t = \arg\min_s \{(1/2)\|s(\mathbf{B}_t^{\mathcal{M}})\|^2 + \mathbb{E}\left[\operatorname{div}(s)(\mathbf{B}_t^{\mathcal{M}})\right]\}.}$$

  - Use **approximation** of $\nabla \log p_{t|0}$ (Varadhan approximation and series expansion).

$$\boxed{\nabla \log p_t = \arg\min_s \{\mathbb{E}\left[\|s(\mathbf{B}_t^{\mathcal{M}}) - \nabla \log p_{t|0}(\mathbf{B}_t^{\mathcal{M}}|\mathbf{B}_0^{\mathcal{M}})\|^2\right]\}.}$$

Euclidean VS compact Riemannian

- **Riemannian score-based generative modeling** (RSGM)
  - Sample from the **forward dynamics**.
  - Train the **score network**.
  - Sample from the **backward dynamics** (initialized at the uniform distribution).
- Differences between the **Euclidean setting** and the **compact manifold setting**.

Ingredient \ Space	Euclidean	Compact manifold
Forward process	Ornstein–Ulhenbeck	Brownian motion
Easy-to-sample distribution	Gaussian	Uniform
Time reversal	(Cattiaux et al., 2021)	This paper
Sampling of the forward process	Direct	Geodesic Random Walk
Sampling of the backward process	Euler–Maruyama	Geodesic Random Walk

Table 1: Differences between SGM on Euclidean spaces and RSGM on compact Riemannian manifolds.

Extension to Schrödinger bridges

- We can extend the **Schrödinger bridge** framework to the manifold setting.
- Difficulty: considering an equivalent of the **mean-matching** technique on manifold (divergence form).

Implicit mean-matching loss

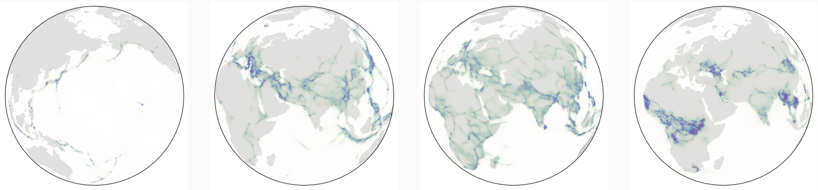
Let $(\mathbf{X}_t)_{t \in [0,T]}$ be a $\mathcal{M}$-valued process with distribution $\mathbb{P} \in \mathcal{P}(\mathrm{C}([0,T], \mathcal{M}))$ such that for any $t \in [0,T]$, $\mathbf{X}_t$ admits a positive density $p_t \in \mathrm{C}^\infty(\mathcal{M})$ w.r.t. p_{ref}. Let $s : [0,T] \to \mathcal{X}\mathcal{M}$. For any $t \in [0,T]$ and $x \in \mathcal{M}$, let

$$\boxed{b(t,x) = -f(t,x) + \mathrm{g}(t,\mathbf{X}_t)^2 \nabla \log p_t(x).}$$

Then, for any $t \in [0,T]$, we have that

$$\boxed{b(t,\cdot) = \arg\min_r \{\mathbb{E}[\tfrac{1}{2}\|f(t,\mathbf{X}_t) + r(\mathbf{X}_t)\|^2 + \mathrm{g}(t,\mathbf{X}_t)^2 \mathrm{div}(r)(\mathbf{X}_t)]\}.}$$

Application



Learned density on Volcano/Earthquake/Flood/Fire datasets.

	Earthquake	Flood	Fire
Mixture of Kent	$0.33_{\pm 0.05}$	$0.73_{\pm 0.07}$	$-1.18_{\pm 0.06}$
Riemannian CNF	$0.19_{\pm 0.04}$	$0.90_{\pm 0.03}$	$-0.66_{\pm 0.05}$
Moser Flow	$-0.09_{\pm 0.02}$	$0.62_{\pm 0.04}$	$-1.03_{\pm 0.03}$
Stereographic Score-Based	$-0.04_{\pm 0.11}$	$1.31_{\pm 0.16}$	$0.28_{\pm 0.20}$
Riemannian Score-Based	$\mathbf{-0.21}_{\pm 0.03}$	$\mathbf{0.52}_{\pm 0.02}$	$\mathbf{-1.24}_{\pm 0.07}$
Dataset size	6120	4875	12809

Table 2: Negative log-likelihood scores for each method on the earth and climate science datasets. Bold indicates best results (up to statistical significance). Means and standard deviations are computed over 5 different runs.

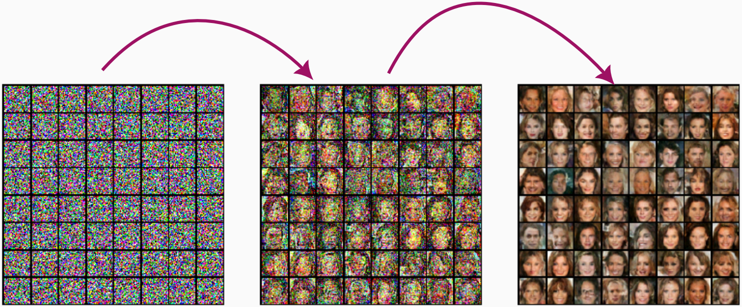
Why generative modeling? (1/2)

- Application in **meteorology**: Ravuri et al. (2021).
  - Prediction of rain in the next 2 hours: **nowcasting**.
  - Solving physical PDEs: **planet scale** predictions days ahead.
  - Struggle for **high resolution** predictions on short time ranges.

- Access to a lot of high quality data: **conditional GAN**.

Image extracted from Ravuri et al. (2021).

Some visual results



Dataset interpolation