

Adaptive denoising diffusion modelling via random time reversal

MFO Mini-Workshop on *Probabilistic Perspectives in Neural Network-Based Machine Learning*

Lukas Trottner

based on joint work with [Sören Christensen](#), [Jan Kallsen](#) and [Claudia Strauch](#)

29 October 2025

University of Stuttgart

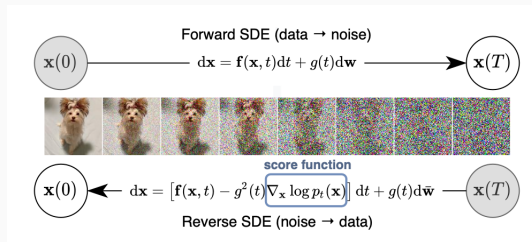
[Kiel University](#) [Heidelberg University](#)



University of Stuttgart
Germany

Denoising diffusion models

- provide an **iterative generative algorithm** to create new samples that approximately match the target distribution p_0 , given a finite number of samples corresponding to an unknown p_0
- general idea: find a **stochastic process** that perturbs p_0 to a new distribution p_T such that
 - 1) p_T or a good approximation thereof is **easy to sample from**, and
 - 2) the perturbation is **reversible** in the sense that we know how to **simulate the time-reversed process**



Source: Song et al. (2021). Score based generative modeling through stochastic differential equations. *ICLR*.

Denoising Diffusion Models

- for some fixed time $T > 0$ consider the **forward model**

$$dX_t = b(t, X_t) dt + \sigma(t, X_t) dW_t, \quad t \in [0, T], X_0 \sim p_0$$

- under sufficient regularity conditions, the forward model has a solution $X = (X_t)_{t \in [0, T]}$ with marginal densities $p_t(x) = \int p_{0,t}(y, x) p_0(dy)$ such that the **time reversal** $\tilde{X}_t = X_{T-t}$ solves

$$d\tilde{X}_t = -\bar{b}(T-t, \tilde{X}_t) dt + \sigma(T-t, \tilde{X}_t) d\bar{W}_t, \quad t \in [0, T], \tilde{X}_0 \sim p_T,$$

where

$$\begin{aligned} \bar{b}_i(t, x) &= b_i(t, x) - \frac{1}{p_t(x)} \sum_{j,k=1}^d \frac{\partial}{\partial x_j} (p_t(x) \sigma_{ik}(t, x) \sigma_{jk}(t, x)) \\ &= b_i(t, x) - (\nabla \cdot \Sigma(t, x))_i - (\nabla \log p_t(x))_i, \quad i = 1, \dots, d, \Sigma = \sigma \sigma^\top \end{aligned}$$

- ~> time-reversed process solves a **time-inhomogeneous SDE**, now with drift $-\bar{b}(T - \cdot, \cdot)$ involving the **score** $\nabla \log p_t$, which depends on the **unknown** data distribution p_0
- ~> score needs to be estimated from the data

Denoising score matching

- denoising score matching:

$$\begin{aligned}\nabla \log p_t(x) &= \frac{\int \nabla_x p_{0,t}(y, x) p_0(dy)}{p_t(x)} = \int \nabla_x \log p_{0,t}(y, x) \underbrace{\frac{p_{0,t}(y, x) p_0(dy)}{p_t(x)}}_{=P(X_0 \in dy | X_t = x)} \\ &= \mathbb{E}[\nabla_2 \log p_{0,t}(X_0, X_t) \mid X_t = x]\end{aligned}$$

and thus

$$\nabla \log p_t \in \arg \min_{s \text{ meas.}} \mathbb{E}[\|s(X_t) - \nabla_2 \log p_{0,t}(X_0, X_t)\|^2]$$

\leadsto given data $(X_0^i)_{i \in [n]} \stackrel{\text{iid}}{\sim} p_0$ define the **denoising score estimator**

$$\hat{s} \in \arg \min_{s \in \mathcal{S}} \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{X_0^i} \left[\int_{\underline{T}}^T \|s(t, X_t) - \nabla_2 \log p_{0,t}(X_0, X_t)\|^2 dt \right],$$

where $0 < \underline{T} \ll T$ and \mathcal{S} is an approximating function class, e.g. **space-time neural networks**

Generative process

On $[0, T - \underline{T}]$, simulate

$$dY_t = \left(-b(T-t, Y_t) + \nabla \cdot \Sigma(T-t, Y_t) + \Sigma(T-t, Y_t) \hat{\mathfrak{g}}(T-t, Y_t) \right) dt + \sigma(T-t, Y_t) dW_t, \quad \mathbb{P}^{Y_0}(dy) \approx p_T(y) dy$$

Output:

$$Y_{T-\underline{T}} \stackrel{d}{\approx} \tilde{X}_{T-\underline{T}} = X_{\underline{T}} \stackrel{d}{\approx} X_0$$

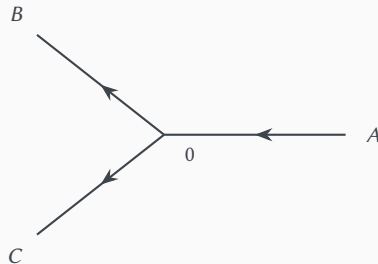
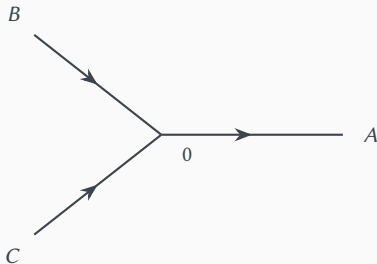
Basic observations

- time reversal at **deterministic** time T forces the backward process to be time-inhomogeneous
- if p_0 has **low-dimensional support** \mathcal{M} , for small t and x close to \mathcal{M} , $\nabla \log p_t(x)$ is approximately **orthogonal** to \mathcal{M} (Stanczuk et al., 2024)¹
- initialising the generative process in a distribution that is not close to \mathbb{P}^{X_T} and simulating for $T - \underline{T}$ time units will not give useful results \leadsto algorithm is **not adaptive** to the noise level in the data

¹Stanczuk et al. (2024). Your diffusion model secretly knows the dimension of the data manifold. *ICML*.

Homogeneous time reversal

- Markov property: “the past and future of a Markov process are conditionally independent given the present” \leadsto time-reversed Markov processes are Markov
- to ensure that a **homogeneous Markov process** remains homogeneous under time reversal, we need to reverse at a suitable **random (life)time ζ** . This can be
 - a randomised stopping time such as an **independent exponential time**;
 - a **last exit time**;
 - a **first hitting time**;
 - any **terminal time**, that is, any stopping time T such that $T = t + T \circ \theta_t$ on $\{T > t\}$
- retaining the strong Markov property under time reversal is a bit more tricky:



h -transforms and time reversal

h -transform

For a possibly killed, homogeneous strong Markov process X with state space S , let h be an excessive function, that is

$$\mathbb{E}_x[h(X_t)] \leq h(x) \quad \text{and} \quad \lim_{t \rightarrow 0} \mathbb{E}_x[h(X_t)] = h(x).$$

Then,

$$p_t^h f(x) = \mathbb{E}_x \left[\frac{h(X_t)}{h(x)} f(X_t) \mathbf{1}_{\{X_t \in S\}} \right] \mathbf{1}_{(0, \infty)}(h(x)), \quad f \in \mathcal{B}_b(\mathbb{R}^d),$$

defines a sub-Markov semigroup. The corresponding Markov process X^h is strong Markov and is called h -transform of X .

- suppose that X is a continuous and [self-dual](#) Feller process (i.e., its generator satisfies $A = A^*$)
- if X^h has a finite killing time ζ , then the time-reversed process $X_t^h = X_{\zeta-t}^h$ is [homogeneous, strong Markov](#) and is a \bar{h} -transform of X .

h -transforming a killed diffusion

- consider a **symmetric** diffusion process

$$dX_t = b(X_t) dt + \sigma(X_t) dW_t$$

with invariant measure m and let Z be its version **killed at an independent exponential time** with parameter $r > 0$

- as an excessive function for Z use

$$h(x) = \int G_r(x, y) \kappa(dy)$$

for the **Green kernel** $G_r(x, y) = \int_0^\infty e^{-rt} p_t(x, y) dy$ and a **representing measure** κ

- $\kappa(dy) = r dy \rightsquigarrow h = 1$ and $Z^h = Z$
- $\kappa(dy) = \frac{1}{G_r(x_0, y)} \beta(dy) \rightsquigarrow Z$ conditioned to have distribution β before killing if started in x_0
- Z is a killed Brownian motion and $\kappa(dy) = \sigma_R(dy)$ for the surface measure σ_R of an R -sphere $\mathbb{S}^{d-1}(R) \rightsquigarrow Z^h$ is killed at last exit from $\mathbb{S}^{d-1}(R)$

A time-homogeneous generative process

Proposition

1. Z^h is an Itô-diffusion with dynamics

$$dZ_t^h = (b(Z_t^h) + \Sigma(X_t) \nabla \log h(X_t)) dt + \sigma(Z_t^h) dW_t$$

outside $\text{supp } \kappa$ and its distribution at the lifetime is given by

$$\mathbb{P}_x(Z_{\zeta-}^h \in dy) = \frac{G_r(x, y)}{h(x)} \kappa(dy)$$

2. Let $\alpha = \mathbb{P}^{Z_0^h}$. Then \tilde{Z}_t^h is an \tilde{h} -transform of Z with initial distribution $\mathbb{P}_\alpha(Z_{\zeta-}^h \in dy)$ and

$$\tilde{h}(x) := \int \frac{G_r(x, y)}{h(y)} \alpha(dy).$$

In particular, \tilde{Z}^h has dynamics

$$d\tilde{Z}_t^h = (b(\tilde{Z}_t^h) + \Sigma(\tilde{Z}_t^h) \nabla \log \tilde{h}(\tilde{Z}_t^h)) dt + \sigma(\tilde{Z}_t^h) d\bar{W}_t,$$

outside $\text{supp } \alpha =: \mathcal{M}$ and $\mathbb{P}_\alpha(\tilde{Z}_{\zeta-}^h \in dy \mid \tilde{Z}_0^h = x) = \frac{G_r(x, y)}{\tilde{h}(x)h(y)} \alpha(dy)$ for $\mathbb{P}_\alpha(Z_{\zeta-}^h \in \cdot)$ -a.e. x .

A time-homogeneous generative process

Idealised algorithm:

1. Initialise $Z_0^{\vec{h}} \sim \tilde{\beta} \approx \mathbb{P}_\alpha(Z_{\zeta-}^h)$
 - for ergodic forward process with stationary distribution μ and small exponential killing rate $r > 0$, choose $\tilde{\beta} = \mu$ [\leftrightarrow [ergodic diffusion model](#)]
 - for exponentially killed BM with small killing rate $r > 0$, choose $\tilde{\beta} = \text{Laplace}(0, (2r)^{-1/2} \mathbb{I}_d)$ [\leftrightarrow [variance exploding diffusion model](#)]
 - for $\kappa(dy) = \frac{1}{G_r(x_0, y)} \delta_z$, choose $\tilde{\beta} = \delta_z$
2. Simulate diffusion $Z^{\vec{h}}$ until killing time and output $Z_{\zeta-}^{\vec{h}}$

Requirements for implementation

1. learn $\nabla \log \vec{h}$ (only a function in space – no time component);
2. learn killing time ζ of $Z^{\vec{h}}$

Learning to kill

Polarity hypothesis

Assume that $\mathcal{M} = \text{supp } \alpha$ is **polar** for X , i.e., for any $x \in \mathbb{R}^d$, $\mathbb{P}_x(\inf\{t > 0 : X_t \in \mathcal{M}\} < \infty) = 0$.

Theorem

Under the polarity hypothesis, the backward process $Z^{\leftarrow h}$ is **killed at first entrance into \mathcal{M}** .

Possible strategies to estimate a δ -fattening $\mathcal{M}_\delta = \{x : \text{dist}(x, \mathcal{M}) \leq \delta\}$ given data $X^1, \dots, X^n \stackrel{\text{iid}}{\sim} \alpha$ and an estimator $\hat{\mathfrak{z}}$ of $\mathfrak{z} := \nabla \log \bar{h}$:

- plug-in approach: estimate \mathcal{M}_δ directly or indirectly by setting $\widehat{\mathcal{M}}_\delta = (\widehat{\mathcal{M}})_\delta$; then set $\hat{\zeta} := \inf\{t \geq 0 : Z_t^{\hat{\mathfrak{z}}} \in \widehat{\mathcal{M}}_\delta\}$
- use explosive behaviour of \mathfrak{z} as $x \rightarrow \mathcal{M}$:

Theorem

Suppose that \mathcal{M} is polar for X and Y solving $dY_t = \sigma(Y_t) dB_t$. Then, it a.s. holds that

$$\zeta = \inf\left\{t \geq 0 : \sup_{s \leq t} |\mathfrak{z}(Z_s^{\leftarrow h})| = \infty\right\} = \inf\left\{t \geq 0 : \|\mathfrak{z}(Z^{\leftarrow h})\|_{L^2([0, t])} = \infty\right\}.$$

Denoising score matching

- for $\mathbb{P}_\alpha(Z_{\zeta_-}^h \in \cdot)$ -a.e. x

$$\begin{aligned}\mathfrak{s}(x) &= \nabla \log \tilde{h}(x) = \frac{1}{\tilde{h}(x)} \int \nabla_x G_r(x, y) \frac{1}{h(y)} \alpha(dy) = \int \nabla_x \log G_r(x, y) \frac{G_r(x, y)}{\tilde{h}(x)h(y)} \alpha(dy) \\ &= \mathbb{E}[\nabla_x \log G_r(x, Z_{\zeta_-}^h) \mid Z_0^h = x] \\ &= \mathbb{E}_\alpha[\nabla_x \log G_r(x, Z_0^h) \mid Z_{\zeta_-}^h = x]\end{aligned}$$

- this implies that on $\mathbb{R}^d \setminus \mathcal{M}_\delta$, \mathfrak{s} agrees $\mathbb{P}_\alpha(Z_{\zeta_-}^h \in \cdot)$ -a.e. with the minimiser of

$$\mathcal{B}(\mathbb{R}^d; \mathbb{R}^d) \ni s \mapsto \mathbb{E}_\alpha \left[\|s(Z_{\zeta_-}^h) - \nabla \log G_r(Z_0^h, Z_{\zeta_-}^h)\|^2 \mathbf{1}_{\{\|Z_{\zeta_-}^h - Z_0^h\| > \delta\}} \right]$$

- note that if $Z^h = Z$, then $\zeta \sim \text{Exp}(r)$ independent of X , $Z_{\zeta_-} = X_\zeta$ has full support and we have

$$\mathbb{E}_\alpha \left[\|s(Z_{\zeta_-}^h) - \nabla \log G_r(Z_0^h, Z_{\zeta_-}^h)\|^2 \mathbf{1}_{\{\|Z_{\zeta_-}^h - Z_0^h\| > \delta\}} \right] = r \mathbb{E}_\alpha \left[\int_0^\zeta \|s(Z_t^h) - \nabla \log G_r(Z_0^h, Z_t^h)\|^2 \mathbf{1}_{\{\|Z_t^h - Z_0^h\| > \delta\}} dt \right]$$

Projection learning

- we don't have to start the backward process approximately in $\mathbb{P}_\alpha(Z_{\zeta_-}^h \in dy)$: it will always be killed on the data support \mathcal{M} and different initialisations will yield different output distributions supported on \mathcal{M} \rightsquigarrow **natural conditioning**
- a natural question is therefore what happens if we don't start the generative process from pure noise but something more informative, say a **masked** or **moderately noised** picture



- it turns out that the natural conditioning aspect entails a **blessing of dimensionality**

Projection learning

Let Z be an **exponentially killed Brownian motion**. Then,

$$\tilde{h}(x) = \int G_r(x, y) \alpha(dy), \quad G_r(x, y) = 2(2\pi)^{-d/2} r \left(\frac{\sqrt{2r}}{|x - y|} \right)^{\frac{d-2}{2}} K_{\frac{d-2}{2}} \left(\frac{\sqrt{2r}}{|x - y|} \right).$$

For large d ,

$$\nabla \log \tilde{h}(x) \approx d \frac{\int \frac{x-y}{|x-y|^d} \alpha(dy)}{\int |x-y|^{2-d} \alpha(dy)}$$

and thus, if there is a unique projection $x^* \in \arg \min_{y \in \mathcal{M}} |x - y|$ of x onto \mathcal{M} , then

$$\nabla \log \tilde{h}(x) \approx d \frac{x^* - x}{|x^* - x|^2} = d \frac{\text{sign}(x^* - x)}{|x^* - x|}$$

Theorem

Let $\delta, \varepsilon > 0$ and fix an observation $x \in \mathbb{R}^d$. If $\alpha(B(x, r)) > \varepsilon$ for some ball $B(x, r)$ with radius $r > 0$ around y , then

$$\mathbb{P}\left(Z_{\zeta-}^{\tilde{h}} \in \mathcal{M} \cap B(x, (1+\delta)r) \mid Z_0^{\tilde{h}} = x\right) \geq 1 - \frac{1}{\varepsilon} (1+\delta)^{2-d}.$$

Projection learning

Consider now estimators $\hat{\mathfrak{s}}_n$, an independent Brownian motion W and let $\widehat{Z}^{\hat{\mathfrak{s}}_n}$ be the process solving

$$d\widehat{Z}_t^{\hat{\mathfrak{s}}_n} = \hat{\mathfrak{s}}_n(\widehat{Z}_t^{\hat{\mathfrak{s}}_n}) \mathbf{1}_{\{t \leq \hat{\zeta}\}} dt + \mathbf{1}_{\{t \leq \hat{\zeta}\}} dW_t, \quad \hat{\zeta} := \inf \left\{ t \geq 0 : \|\widehat{Z}^{\hat{\mathfrak{s}}_n}\|_{L^2[0,t]} > M \right\}.$$

Theorem

Fix an observation $x \in \mathbb{R}^d$. Suppose that

- for any $\tilde{\delta}, \delta, \varepsilon > 0$ it holds for sufficiently large n that

$$\mathbb{P} \left(\left\| (\hat{\mathfrak{s}}_n(Z^{\tilde{h}}) - \mathfrak{s}(Z^{\tilde{h}})) \mathbf{1}_{\{Z^{\tilde{h}} \notin \mathcal{M}_{\tilde{\delta}}\}} \right\|_{L^2(\zeta)} > \delta \mid Z_0^{\tilde{h}} = x \right) < \varepsilon$$

- for any $n \in \mathbb{N}$ and $\tilde{\delta} > 0$, the function $\hat{\mathfrak{s}}_n$ is $L_{\tilde{\delta}}$ -Lipschitz on $\mathcal{M}_{\tilde{\delta}}^c$

Let $\delta, \varepsilon, \tilde{\delta}, \tilde{\varepsilon} > 0$. If $\alpha(B(x, r)) > \varepsilon$, then, for sufficiently large $M > 0$ and $n \in \mathbb{N}$,

$$\mathbb{P}(\widehat{Z}_{\hat{\zeta}}^{\hat{\mathfrak{s}}_n} \in \mathcal{M}_{\tilde{\delta}} \cap B(x, (1 + \delta)r) \mid \widehat{Z}_0^{\hat{\mathfrak{s}}_n} = x) > 1 - \frac{1}{\varepsilon}(1 + \delta)^{2-d} - \tilde{\varepsilon}.$$

Thank you for your attention!