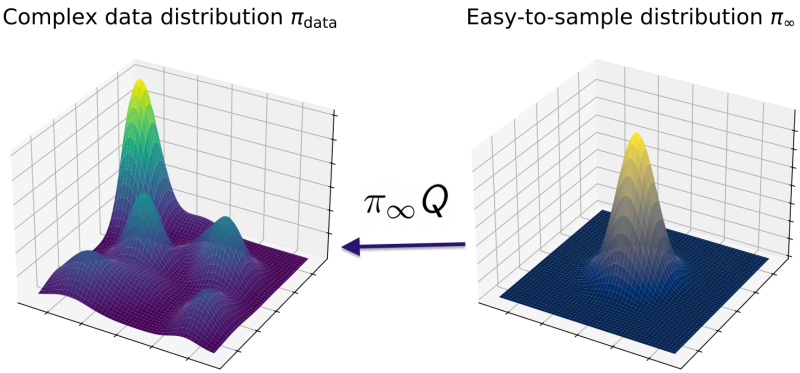
An Analysis of Noise Schedules in Score-Based Generative Modeling Algorithms

Stanislas Strasman, Antonio Ocello, Claire Boyer, Sylvain Le Corff, Vincent Lemaire

SCIENCES SORBONNE UNIVERSITÉ

Generative modeling framework

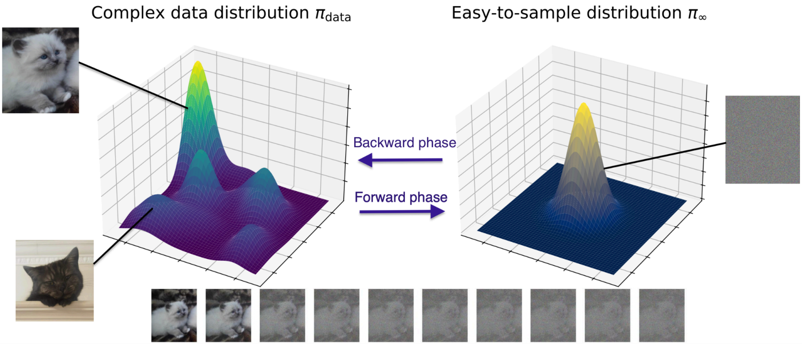
- $\mathcal{D} = \{x_i\}_{i=1}^n \in (\mathbb{R}^d)^n$ a collection of i.i.d. samples from an unknown distribution π_{data}[1].
- Goal: **generate new samples from** π_{data} (i.e. find a proba π_∞ and a simulable kernel Q such that $\pi_{\text{data}} \simeq \pi_\infty Q$).



[1] In this presentation, π will be used interchangeably to denote a probability distribution and its associated probability density function (p.d.f.).

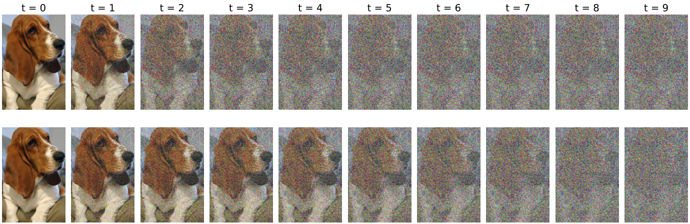
SGMs Philosophy

- "Creating noise from data is easy; creating data from noise is generative modeling." (Song et al., 2021)



What is the appropriate amount of noise ?

- The noising/denoising process is at the core of SGMs.
- SGMs require to **hand-design** the intensity and the form of the noising procedure.
- **Little is known theoretically**, we only know **best practices** from experience and empirical studies (Nichol and Dhariwal, 2021; Guo et al., 2023; Chen, 2023).



Can we tell which is better?

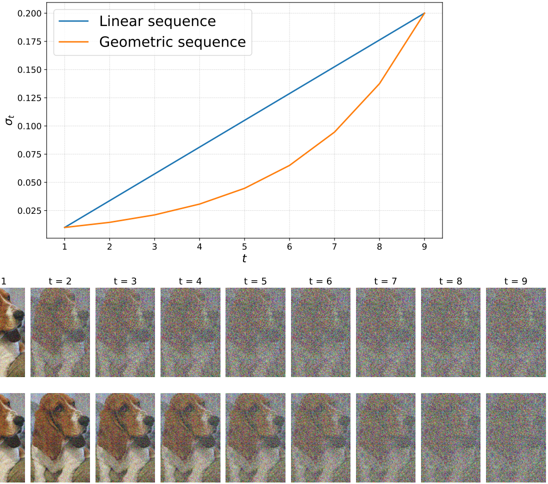


Figure: $X_t = X_0 + \sigma_t \cdot Z$, with $X_0 \sim p_{data}$ and $Z \sim \mathcal{N}(0, I_d)$. For a sequence of positive scalar σ_t for $t \in \{1, ..., 8\}$ with $\sigma_1 = 0.01$ and $\sigma_8 = 0.2$.

Table of Contents

Table of Contents

Foundational insights from DDPM

- Denoising Diffusion Probabilistic Models (Sohl-Dickstein et al., 2015; Ho et al., 2020).
- Consider the **Markov chain** [2] $X_0, X_1, X_2 ... \in \mathbb{R}^d$ starting at $X_0 \sim \pi_{\mathrm{data}}$ and run $N \in \mathbb{N}$ times until the distribution of X_N is close to an *easy-to-sample* prior π_∞:

$$\pi_N(x_N) = \int \pi(x_0, x_1, ..x_N)\mathrm{d}x_0 ... \mathrm{d}x_{N-1} \approx \pi_\infty \,.$$

- By the **Markov property** and Bayes' formula,

$$\begin{aligned} \pi(x_0, x_1, ..x_N) &= \pi_{\mathrm{data}}(x_0) \prod_{k=1}^{N} \pi_{k|k-1}(x_k|x_{k-1}) && \text{(Forward)} \\ &= \pi_N(x_N) \prod_{k=1}^{N} \pi_{k-1|k}(x_{k-1}|x_k) && \text{(Backward)} \end{aligned}$$

[2](admitting positive transition probabilities)

DDPM forward main intuitions

- **Forward phase**: hand-designed Gaussian transition kernels such that for $k \in \{1, 2, \dots, N\}$,

$$\pi_{k|k-1}(x_k|x_{k-1}) = \mathcal{N}(x_k; \sqrt{1-\beta_k}\, x_{k-1}, \beta_k I_d)\,,$$

  with noise scale $0 < \beta_1 \leq \beta_2 \leq \cdots \leq \beta_N < 1$.
- The log-density of the transition kernel is

$$\log \pi_{k|k-1}(x_k|x_{k-1}) \propto -\frac{1}{2\beta_k} \left\| x_k - \sqrt{1-\beta_k}\, x_{k-1} \right\|^2 .$$

- Assume that the β_k **are small enough** such that

$$\log \pi_k(.) = \log \pi_{k-1}(.) + \mathcal{O}(\beta_k)\,,$$

  and

$$\left\| x_k - \sqrt{1-\beta_k}\, x_{k-1} \right\|^2 = \|x_k - x_{k-1}\|^2 + o(\beta_k)\,.$$

DDPM backward steps remain Gaussian !

- Using Bayes' Rule,

$$\begin{aligned}\log \pi_{k-1|k}(x_{k-1}|x_k) &= \log \frac{\pi_{k|k-1}(x_k|x_{k-1})\,\pi_{k-1}(x_{k-1})}{\pi_k(x_k)} \\ &\propto \log \pi_{k|k-1}(x_k|x_{k-1}) + \log \pi_{k-1}(x_{k-1}) \\ &\propto -\frac{1}{2\beta_k}\|x_k - x_{k-1}\|^2 + \log \pi_k(x_{k-1}) + \mathcal{O}(\beta_k)\,.\end{aligned}$$

- Using Taylor Expansion,

$$\log \pi_k(x_{k-1}) = \log \pi_k(x_k) + (x_{k-1} - x_k)^\top \nabla \log \pi_k(x_k) + \mathcal{O}(\|x_{k-1} - x_k\|^2)\,,$$

  with $\|x_{k-1} - x_k\|^2 \sim \mathcal{O}(\beta_k)$.

- Completing the square, and neglecting terms of order β_k, the **conditional backward is Gaussian**:

$$\log \pi_{k-1|k}(x_{k-1}|x_k) \propto -\frac{1}{2\beta_k}\left\| x_{k-1} - \underbrace{(x_k + \beta_k \nabla \log \pi_k(x_k))}_{\mu_k} \right\|^2 .$$

DDPM reverse process and sampling

- When the β_k are small the **reverse conditional distributions** $\pi_{k-1|k}(x_{k-1} \mid x_k)$ are **approximately Gaussian**.
- This is a **score approximation problem** or **denoising problem**. In particular, in particular one might see the connection with Tweedie's formula (Robbins, 1956),

$$\mu_k = \mathbb{E}[x_{k-1} \mid x_k] = x_k + \beta_k \underbrace{\nabla \log \pi_k(x_k)}_{\text{score function}}.$$

- To estimate the score, one can train a **neural network**[3]

$$s_\theta(x_k, k) : \{1, 2, .., N\} \times \mathbb{R}^d \to \mathbb{R}^d$$

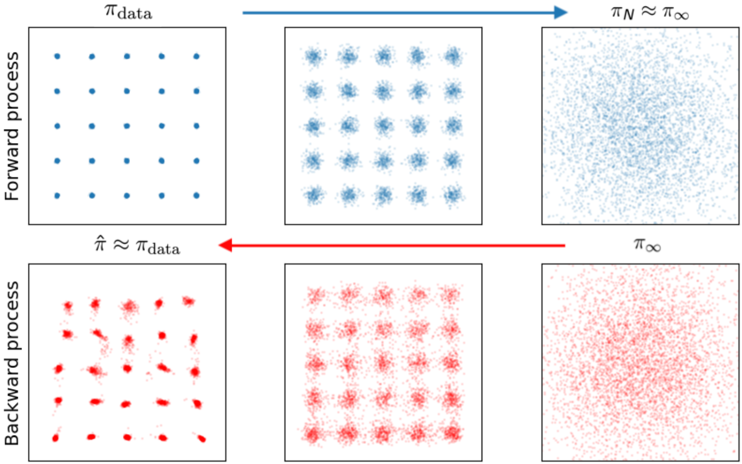
- To **sample from the reverse process**, sample from $\pi_\infty \approx \pi_N$ and apply **ancestral sampling** on the approximated backward transitions.

[3] (no details given at this point).

Table of Contents

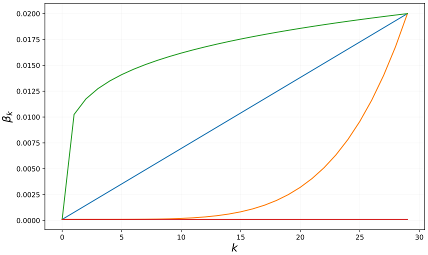
DDPM on Gaussian mixture model

- DDPM trained on a **2-dimensional mixture of 25 Gaussian random variables.**
- The resulting diffusion process is given below on a batch of **1000 samples**.

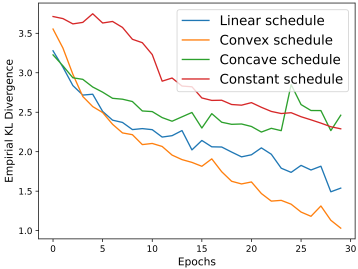


Impact of the noise schedule on the generation quality

- In **low dimension** the KL-divergence can be estimated **using histograms**, several schedules are tested.



Noise schedule



Empirical $\mathrm{KL}(\pi_{\mathrm{data}}|\hat{\pi})$.

✓ The noise schedule does seem to impact the generation quality.

Table of Contents

Leveraging the power of continuous-time analysis

- **Convergence results** for diffusion models are established in a **continuous setting** leveraging stochastic calculus tools.
- Indeed, let $0 \leq \Delta \leq 2\Delta \leq .. \leq N\Delta = T$ with $\Delta = T/N$ and set for all $k \in [1, N]$, $\beta_k = 2\Delta$. When $N \to \infty$:

$$\begin{aligned} X_{k+\Delta} &= \sqrt{1 - 2\Delta} X_k + \sqrt{2\Delta} Z \\ &\approx_{\Delta \to 0} (1 - \Delta) X_k + \sqrt{2\Delta} Z \,, \end{aligned}$$

Hence, the **limiting process of DDPM** is, for $t \in [0, T]$,

$$\mathrm{d}X_t = -X_t \,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t \,.$$

SGMs through SDE : forward process

- For some diffusion time-horizon $T > 0$, the **forward process** $\left(\overrightarrow{X}_t\right)_{t\in[0,T]}$ is solution to an Ornstein-Uhlenbeck process:

$$\mathrm{d}\overrightarrow{X}_t = -\overrightarrow{X}_t \mathrm{d}t + \sqrt{2}\mathrm{d}B_t\,, \quad X_0 \sim \pi_{\mathrm{data}}\,.$$

- Let Q_t be the semi-group associated with $\overrightarrow{X}_t$ and let $\pi_t = \pi_{\mathrm{data}} Q_t$.
- In the time limit, the above transports π_{data} to a standard Gaussian distribution π_∞ by progressively adding (Gaussian) noise.

SGMs through SDE: more on the forward process

- As in DDPM the noising procedure implies a scaling down of the of the data points $\mathrm{d}\overrightarrow{X}_t = -\overrightarrow{X}_t \mathrm{d}t$,

SGMs through SDE: more on the forward process

- ... and a Gaussian noising process $\mathrm{d}\overrightarrow{X}_t = \sqrt{2}\mathrm{d}B_t$,

SGMs through SDE: more on the forward process

SGMs through SDE: backward process

- Under mild conditions the forward process admits a **time-reversed process** (Anderson, 1982; Cattiaux et al., 2021), i.e. in law,

$$\left(\overleftarrow{X}_t\right)_{t\in[0,T]} = \left(\overrightarrow{X}_{T-t}\right)_{t\in[0,T]}$$

with,

$$\mathrm{d}\overleftarrow{X}_t = \left(\overleftarrow{X}_t + 2\underbrace{\nabla \log \pi_{T-t}\left(\overleftarrow{X}_t\right)}_{\text{score function}}\right)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t\,, \quad \overleftarrow{X}_0 \sim \pi_T\,.$$

- The score term will drive the backward equation in **regions of space of high probability**.
- This gives a natural way to construct a **backward process** and therefore a generative model as in DPPM.

A variety of time-homogeneous convergence results

- Using this framework a **variety of upper bounds** to the **distance between the data distribution and the generated distribution** $d\left(\pi_{\text{data}}, \hat{\pi}\right)$ have been established for various metrics:
  - For the total variation distance: De Bortoli et al. (2021).
  - For the Kullback-Leibler divergence: Conforti et al. (2023); Bortoli et al. (2023); Chen et al. (2023); Chen (2023).
  - For the Wasserstein distance: Lee et al. (2022, 2023); Bruno et al. (2023); Gao et al. (2023).
- **Remark:** one can convert KL bounds into total variation bounds using Pinsker's inequality:

$$\|\pi_{\text{data}} - \hat{\pi}\|_{\text{TV}} \leq \sqrt{\frac{1}{2}\mathrm{KL}\left(\pi_{\text{data}}\|\widehat{\pi}\right)}\,.$$

Table of Contents

Adapted theoretical framework: time-inhomogeneous SDE

⚠ An homogeneous forward implies a specific noise schedule choice.

1. **Forward process** now depends on $\beta : [0, T] \mapsto \mathbb{R}_{>0}$,

$$\mathrm{d}\overrightarrow{X}_t = -\frac{\beta(t)}{2\sigma^2}\overrightarrow{X}_t \mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}B_t\,, \quad \overrightarrow{X}_0 \sim \pi_{\mathrm{data}}\,.$$

2. **Backward process**,

$$\mathrm{d}\overleftarrow{X}_t = \left(\frac{\beta(T-t)}{2\sigma^2}\overleftarrow{X}_t + \beta(T-t)\underbrace{\nabla \log \pi_{T-t}\left(\overleftarrow{X}_t\right)}_{\textbf{score function}}\right)\mathrm{d}t + \beta(T-t)\mathrm{d}B_t\,, \quad \overleftarrow{X}_0 \sim \pi_T\,.$$

💡 How to go from this result to a practically viable **generative algorithm**?

SGMs in Practice I: mixing time.

- We let Q_t be the semigroup of $\overleftarrow{X}_t$ defined as

$$Q_t(x, \mathrm{d}y) = \mathbb{P}\left(\overleftarrow{X}_t \in \mathrm{d}y | \overleftarrow{X}_0 = x\right) .$$

- Recall that **time-reversal holds when $\overleftarrow{X}_\mathbf{0} \sim \pi_\mathbf{T}$**, i.e.

$$\pi_{\mathsf{data}} = \pi_T Q_T .$$

- But π_t depends on π_{data}:

$$\pi_t(x_t) = \int_{\mathbb{R}^d} \underbrace{\pi_t(x_t | x_0)}_{\text{p.d.f. of } \overrightarrow{X}_t | X_0} \pi_{\mathsf{data}}(x_0)\, \mathrm{d}x_0 .$$

- In practice, we want a specified and easy-to-sample probability π_∞ to initialize the generative model.

SGMs in Practice I: mixing time.

- 💡 **Idea:** leverage the ergodicity of the O–U kernel.
- **Forward process** admits time marginal with $Z \sim \mathcal{N}(0, I_d)$ and $Z \perp X_0$:

$$\overrightarrow{X}_t = m_t X_0 + \sigma_t Z \,,$$

where:

$$m_t = \exp\left\{ - \int_0^t \frac{\beta(s)}{2\sigma^2} \, \mathrm{d}s \right\}, \qquad \sigma_t^2 = \sigma^2 \left(1 - {m_t}^2\right) \,.$$

- For T large,

$$\pi_T \approx \pi_\infty \sim \mathcal{N}\left(0, \sigma^2 I_d\right) \,.$$

⚠ **Mixing Time Error:** $\pi_{\text{data}} \simeq \pi_\infty Q_T$

SGMs in practice II: learn the score function

- Recall that the backward process depends on the score function $\nabla \log \pi_t(x)$.
- We train a **deep neural network** $s_\theta : [0, T] \times \mathbb{R}^d \mapsto \mathbb{R}^d$ to minimize:

$$\mathcal{L}_{\text{explicit}}(\theta) = \mathbb{E}\left[\left\| s_\theta\left(\tau, \overrightarrow{X}_\tau\right) - \nabla \log \pi_\tau\left(\overrightarrow{X}_\tau\right)\right\|^2\right],$$

  with $\tau \sim \mathcal{U}(0, T)$ independent of the forward process $(\overrightarrow{X}_t)_{t \geq 0}$.
- But $\pi_\tau(x)$ is **unknown !**

SGMs in practice II: learn the score function

- **Idea:** its **conditional version** shares the same optimum (Hyvärinen and Dayan, 2005; Vincent, 2011):

$$\mathcal{L}_{\text{score}}(\theta) = \mathbb{E}\left[\|s_\theta\left(\tau, \overrightarrow{X}_\tau\right) - \nabla \log \pi_\tau\left(\overrightarrow{X}_\tau | X_0\right)\|^2\right].$$

- The conditional score is explicit :

$$\nabla \log \pi_\tau(\overrightarrow{X}_\tau | X_0) = \frac{m_\tau X_0 - \overrightarrow{X}_\tau}{\sigma_\tau^2} = -\frac{Z}{\sigma_\tau}$$

- Score matching Neural Networks writes as,

$$\mathcal{L}_{\text{score}}(\theta) = \mathbb{E}\left[\left\|s_\theta\left(\tau, \overrightarrow{X}_\tau\right) + \frac{Z}{\sigma_\tau}\right\|^2\right].$$

⚠ **Approximation error**: $\pi_{\text{data}} \approx \pi_\infty Q_T^\theta$

SGMs in practice III: simulate from the backward kernel

- Contrary to the forward process the backward is **non-linear**.
- **Idea: discretize** $[0, T]$ by N points with $t_k = kh$ and $h = T/N$, we let $t = t_k$ if $kh \leq t \leq (k+1)h$.
- Consider the **Exponential Integrator scheme**:

$$\mathrm{d}\overleftarrow{X}^{\theta}_{t,N} = \left(\frac{\beta(T-t)}{2\sigma^2}\overleftarrow{X}^{\theta}_{t,N} + \beta(T-t)s_\theta\left(T - t_k \overleftarrow{X}^{\theta}_{t_k,N}\right)\right)\mathrm{d}t + \beta(T-t)\mathrm{d}B_t\,, \overleftarrow{X}_0 \sim \pi_\infty.$$

Discretization error: $\pi_{\mathrm{data}} \approx \pi_\infty Q^{\theta}_{T,N} := \widehat{\pi}^{(\beta,\theta)}_{\infty,N}$

Table of Contents

KL upper bound with minimal hypotheses

Theorem (S. et al 2024)

Hyp: (i) β is continuous, positive, $\nearrow$, with $\int_0^\infty \beta(t)\mathrm{d}t = \infty$
(ii) Novikov's condition on the difference between the actual and estimated score functions.
(iii) $\mathcal{I}(\pi_{\mathrm{data}}|\pi_\infty) < \infty$. Then,

$$\mathrm{KL}\left(\pi_{\mathrm{data}}\|\widehat{\pi}_{\infty,N}^{(\beta,\theta)}\right) \leq \underbrace{\mathrm{KL}\left(\pi_{\mathrm{data}}\|\pi_\infty\right)\exp\left\{-\frac{1}{\sigma^2}\int_0^T \beta(s)\mathrm{d}s\right\}}_{\text{Mixing time}}$$

$$+\underbrace{\sum_{k=0}^{N-1}\mathcal{E}_{\theta,k}^{\beta}\int_{T-t_{k+1}}^{T-t_k}\beta(t)\mathrm{d}t}_{\text{Approx. error}} + \underbrace{2h\beta(T)\mathcal{I}(\pi_{\mathrm{data}}|\pi_\infty)}_{\text{Discr. error}}.$$

with $\mathcal{E}_{\theta,k}^{\beta} = \mathbb{E}\left[\left\|\nabla \log \pi_{T-t_k}\left(\overrightarrow{X}_{T-t_k}\right) - s_\theta\left(T-t_k, \overrightarrow{X}_{T-t_k}\right)\right\|^2\right].$

Sketch of proof

- Time reversal, data processing inequality and Girsanov theorem,

$$
\begin{aligned}
\mathrm{KL}\left(\pi_{\mathrm{data}} \middle\| \pi_\infty Q^{\theta}_{T,N}\right) &= \mathrm{KL}\left(\pi_T Q_T \middle\| \pi_\infty Q^{\theta}_{T,N}\right) \\
&\leq \mathrm{KL}\left(\pi_T \| \pi_\infty\right) + \frac{1}{2}\int_0^T \mathbb{E}\left[\left\|\beta(t)\left(\nabla \log \pi_t\left(\tau_t, \overleftarrow{X}_t\right) - s_\theta(\tau_k, \overleftarrow{X}_{t_k})\right)\right\|^2\right]\mathrm{d}t \\
&\leq \underbrace{\mathrm{KL}\left(\pi_T \| \pi_\infty\right)}_{E_1(\beta)} + \underbrace{\frac{1}{2}\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}} \mathbb{E}\left[\left\|\beta(t)\left(\nabla \log \pi_{\tau_k}\left(\tau_k, \overleftarrow{X}_{\tau_k}\right) - s_\theta(\tau_k, \overleftarrow{X}_{t_k})\right)\right\|^2\right]\mathrm{d}t}_{E_2(\beta,\theta)} \\
&\quad + \underbrace{\frac{1}{2}\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}} \mathbb{E}\left[\left\|\beta(t)\left(\nabla \log \pi_{\tau_t}\left(\tau_t, \overleftarrow{X}_t\right) - \nabla \log \pi_{\tau_k}\left(\tau_k, \overleftarrow{X}_{\tau_k}\right)\right)\right\|^2\right]\mathrm{d}t}_{E_3(\beta)}\,.
\end{aligned}
$$

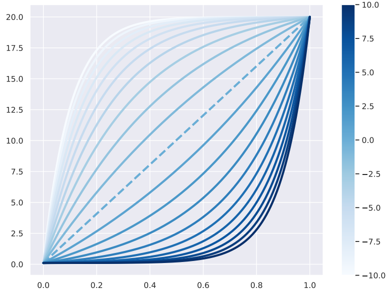
with $\tau_t = T - t$ and $\pi_k = T - t_k$.

Sketch of proof

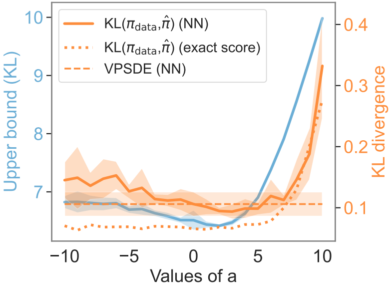
- $E_1(\beta)$ (mixing time error) represents the convergence of the forward process to its stationary distribution π_∞. The rate is given by **Log-Sobolev inequalities**.
- $E_2(\beta, \theta)$ (approximation error) is the quality of the learning process (L_2-error assumed to be finite, i.e. $\mathcal{E}_{\theta,k}^{\beta} \leq \infty$) at every discretization step.
- $E_3(\beta)$ (discretization error) arises from discretizing a continuous-time process into finite steps. Follows from summing up errors between discterization steps and using moments bounds on $\mathbb{E}\left[\left\|\nabla \log \pi_t\left(\overleftarrow{X}_t\right)\right\|^2\right]$ (using time-reversal).

Numerical analysis I

- The effect of $\beta(.)$ is rather complicated to be studied analytically but numerical experiments are possible.



Noise schedules β_a



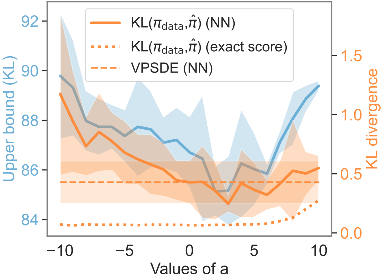
Isotropic Gaussian $\mathcal{N}(\mathbf{1}_d, 0.5\mathrm{I}_d)$

Figure: Comparison of the empirical KL divergence (mean $\pm$ std over 10 runs) between π_{data} and $\widehat{\pi}_{\infty,N}^{(\beta_a,\theta)}$ (orange) and the upper bound (blue) across parameter a for noise schedule β_a, $d = 50$.

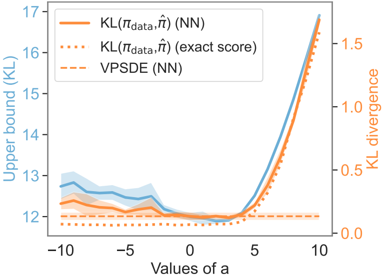
✓ the noise schedule has an impact on the generation quality (rather expected).

✓ the upper bound captures this effect (maybe less expected).

✓ results are in line with heuristics.



Anisotropic $\mathcal{N}\left(\mathbf{1}_d, \Sigma^{(\text{heterosc})}\right)$ [4]



Correlated $\mathcal{N}\left(\mathbf{1}_d, \Sigma^{(\text{corr})}\right)$ [5]

[4] $\Sigma^{(\text{heterosc})}$ is diag. and $\Sigma^{(\text{heterosc})}_{jj} = 1$ for $1 \leq j \leq 5$, and $\Sigma^{(\text{heterosc})}_{jj} = 0.01$ otherwise.

[5] $\Sigma^{(\text{corr})}$ is diag. $\mathbf{1}$ and $\Sigma^{(\text{corr})}_{jj'} = 1/\sqrt{|j - j'|}$ for $1 \leq j \neq j' \leq d$.

Table of Contents

Refined Wasserstein bound

Theorem (S. et al 2024)

Hyp: (i) π_t *is* C_t*-strongly log-concave for* $t \in [0, T]$

(ii) $\nabla \log \pi_t$ *is* L_t*-smooth for* $t \in [0, T]$

(iii) there exists M *such that*

$\sup_{k\in 0,\dots,N-1} \sup_{t_k \le t \le t_{k+1}} |\nabla \log \pi_t - \nabla \log \pi_{t_k}|_{L_2} \le Mh(1+|x|)$

Then,

$$\mathcal{W}_2\left(\pi_{\mathrm{data}}, \widehat{\pi}_{\infty,N}^{(\beta,\theta)}\right) \le \underbrace{\mathcal{W}_2\left(\pi_{\mathrm{data}}, \pi_\infty\right) \exp\left(-\int_0^T \frac{\beta(t)}{\sigma^2}\left(1 + C_t\sigma^2\right)\mathrm{d}t\right)}_{\text{Mixing Time}}$$

$$+ \sum_{k=0}^{N-1}\left(\int_{t_k}^{t_{k+1}} \bar{L}_t \bar{\beta}(t)\mathrm{d}t\right)\left(\frac{\sqrt{2h\beta(T)}}{\sigma} + \frac{h\beta(T)}{2\sigma^2} + \int_{t_k}^{t_{k+1}} 2\bar{L}_t\bar{\beta}(t)\mathrm{d}t\right) B$$

$$+ \mathcal{E}^\beta T\beta(T) + MhT\beta(T)(1+2B)$$

with $B = (\mathbb{E}[\|X_0\|^2] + \sigma^2 d)^{1/2}$, $\bar{L}_t = L_{T-t}$, $\bar{\beta}(t) = \beta(T-t)$, *and*
$\mathcal{E}^\beta = \sup_{k\in\{0,..,N-1\}} \left\| \nabla \log \pi_{T-t_k}\left(\bar{X}_{t_k}^\theta\right) - s_\theta\left(T-t_k, \bar{X}_{t_k}^\theta\right)\right\|_{L_2}$.

Corollary

If $\nabla \log \pi_{\mathrm{data}}$ *is* L_0*-Lipschitz and* $\log \pi_{\mathrm{data}}$ *est* C_0*-strongly concave with* $C_0 > 1/\sigma^2$*, then*

$$\mathcal{W}_2\left(\pi_{\mathrm{data}}, \widehat{\pi}_{\infty,N}^{(\beta,\theta)}\right) \leq \mathcal{W}_2\left(\pi_{\mathrm{data}}, \pi_\infty\right) \exp\left(-\int_0^T \frac{\beta(t)}{\sigma^2}\left(1 + C_t'\sigma^2\right)\mathrm{d}t\right)$$

$$+ \sqrt{h} L_0 \beta(T) T \frac{\sqrt{2\beta(T)}}{\sigma}$$

$$+ h\beta(T) T \left(L_0\left(\frac{1}{2\sigma^2} + 2L_0\right)\beta(T)B + M(1+2B)\right) + \varepsilon T \beta(T).$$

with

$$C_t' = \frac{1}{m_t^2/C_0 + \sigma^2\left(1 - m_t^2\right)} - \frac{1}{\sigma^2},$$

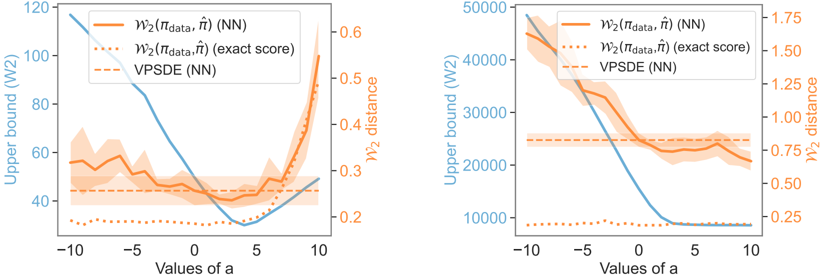
$$m_t = \exp\left(-\frac{1}{2\sigma^2}\int_0^t \beta(s)\mathrm{d}s\right).$$

Sketch of Proof

$$\mathcal{W}_2\left(\pi_{\mathrm{data}}, \widehat{\pi}_{\infty,N}^{(\beta,\theta)}\right) \leq \mathcal{W}_2\left(\pi_{\mathrm{data}}, \pi_\infty Q_T\right) + \mathcal{W}_2\left(\pi_\infty Q_T, \pi_\infty Q_T^{N,\theta}\right)$$

1. **Mixing time error**:
   - Contractivity of the O.U. kernel for the forward process:
   $$\mathcal{W}_2\left(\pi_T, \pi_\infty\right) \leq \mathcal{W}_2\left(\pi_{\mathrm{data}}, \pi_\infty\right) \exp\left(-\int_0^T \frac{\beta(t)}{2\sigma^2}\mathrm{d}t\right) .$$
   - Contractivity of the backward process under strong log-concavity of the score function:
   $$\mathcal{W}_2\left(\pi_T Q_T, \pi_\infty Q_T\right) \leq \mathcal{W}_2\left(\pi_{\mathrm{data}}, \pi_\infty\right) \exp\left(-\int_0^T \frac{\beta(t)}{\sigma^2}\left(1 + C_t\sigma^2\right)\mathrm{d}t\right) .$$
2. **Approximation error and discretization error**:
   - Control the difference between the true backward process $\overleftarrow{X}_t^{\infty}$ and the discretized process $\bar{X}_t^{\theta}$ using the forward backward relationship.

Numerical analysis II



(a) Isotropic $\mathcal{N}(\mathbf{1}_d, 0.5\mathrm{I}_d)$ (b) Correlated $\mathcal{N}\left(\mathbf{1}_d, \Sigma^{(\mathrm{corr})}\right)$

Figure: Comparison of the empirical $\mathcal{W}_2$ distance (mean $\pm$ std over 10 runs) between π_{data} and $\widehat{\pi}^{(\beta,\theta)}_{\infty,N}$ (orange) and the related upper bounds (blue) across parameter a for noise schedule β_a, $d = 50$.

Calibration of σ^2 and Forward Regularization

Assuming the hypotheses of the corollary are verified on π_{data}, the choice of the stationary distribution $\mathcal{N}(0, \sigma^2 I_d)$ is **not obvious** *a priori*.

- If $\sigma^2 \uparrow$, then $L_t \downarrow$; we **"gain in regularity"** of the score function. Also,

$$\frac{1}{\sigma^2} \leq L_0 \implies \forall t \in [0, T], \quad L_t \leq L_0 \, .$$

- If $\sigma^2 \uparrow$, then $C_t \downarrow$; we **"lose in concavity"** of the score function. Also,

$$\frac{1}{\sigma^2} \leq C_0 \implies \forall t \in [0, T], \quad C_t \leq C_0 \, .$$

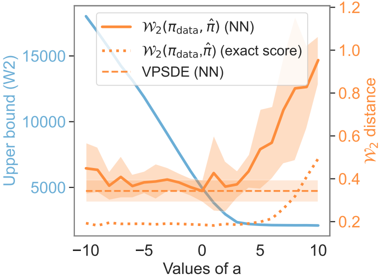
Data conditioning and Forward Regularization

- The **Gaussian experiment** reveals that **data conditioning** is **crucial**.
- If $\pi_{\mathrm{data}} = \mathcal{N}(\mu, \Sigma)$ then:

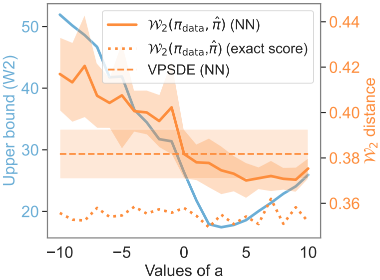
$$C_0 = \frac{1}{\lambda_{\max}(\Sigma)} \quad \text{and} \quad L_0 = \frac{1}{\lambda_{\min}(\Sigma)}.$$

- The smaller the ratio $L_0/C_0 = \lambda_{\max}/\lambda_{\min}$, the tighter the bound, regardless of the choice of σ^2.

Illustration of the Impact of Conditioning



(a) Anisotropic $L_0/C_0 = 100$

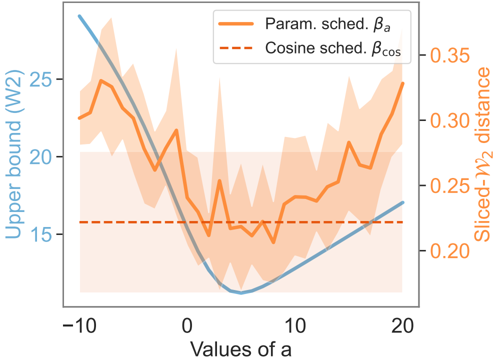


(b) Rescaled Anisotropic $L_0/C_0 = 1$

Figure: Comparison of empirical $\mathcal{W}_2$ distances (mean $\pm$ std over 10 runs) between π_{data} and $\widehat{\pi}^{(\beta,\theta)}_{\infty,N}$ (orange) and the upper bound (blue) for different values of parameter a in the *schedule* β_a, with $d = 50$.

Beyond the Gaussian setting: Funnel distribution

- $\pi_{\mathrm{data}}(x) = \mathcal{N}(x_1; 0, 1) \prod_{j=2}^{d} \mathcal{N}(x_j; 0, \exp(x_1))$ in dimension $d = 50$.
- To evaluate the data generation we use the **2-Sliced Wasserstein** distance.



✓ The Wasserstein bound seems to hold for more general distributions.

Funnel distribution scatter plot

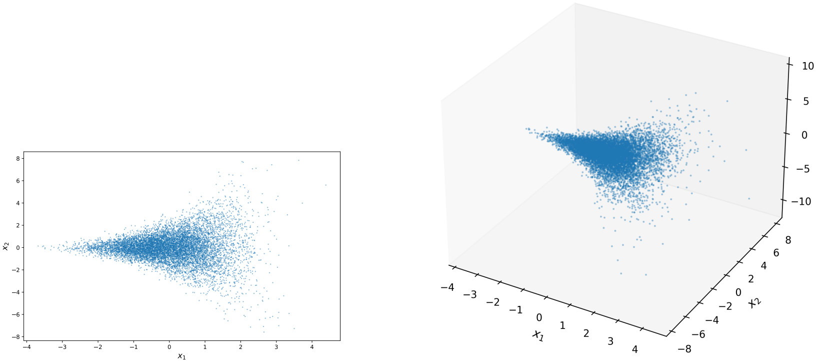
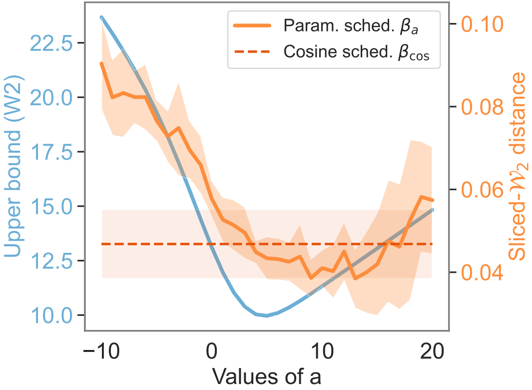


Figure: 10 000 samples from a funnel distribution in dimension 50. Plot of the 1st and 2nd dimension (left) and plot of the 1st, 2nd and 3rd dimension (right).

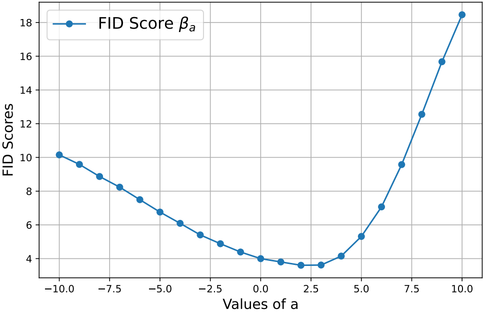
Beyond the Gaussian setting: back to GMM !

- $\pi_{\mathrm{data}}(x) = \frac{1}{25}\sum_{(j,k)\in\{-2,\dots,2\}^2} \varphi_{\mu_{jk},\Sigma_d}(x)$ with $\varphi_{\mu_{jk},\Sigma_d}$ denoting the probability density function of the Gaussian distribution in dimension $d = 50$.



Beyond the Gaussian setting: images ?

- Using **pretrained denoiser nets** from Karras et al. (2022) on CIFAR10 seems to validate the parametric family β_a and is in line with the optimal choices of a^*.



Final word and extension

- Some works consider **homogeneous forward** and **non-homogeneous discretization** steps $\Delta_k = t_{k+1} - t_k$, the Euler-Maruyama updates write:

$$X_{t_{k+1}} = X_{t_k} - X_{t_k}\Delta_k + \sqrt{2\Delta_k}Z_k\,.$$

- One can retrieve an **inhomogeneous SDE with constant discretization** steps setting $\Delta = T/N$ and

$$\beta(t_k) = \frac{2\Delta_k}{\Delta}\,.$$

- However, this approach will prescribe noise schedule choice only at the disretization points which is somehow less informative.
- The previous upper bounds assumed constant step size only for the sake of clarity.

Final word and extension

- We saw theoretically and empirically that the **noise schedule has an impact in the generation quality for SGMs**.
- This line of work pave the way for **noise schedule optimization** dependent on the data properties and on the other hyperparameters (discretization steps, stationnary law, diffusion time).
- However the estimation of the bound remains **tricky in high dimension** due to error terms difficult to estimate.

References I

B. D. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

V. D. Bortoli, J. Thornton, J. Heng, and A. Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling, 2023.

S. Bruno, Y. Zhang, D.-Y. Lim, Ö. D. Akyildiz, and S. Sabanis. On diffusion-based generative models and their error bounds: The log-concave case with full convergence estimates. *arXiv preprint arXiv:2311.13584*, 2023.

P. Cattiaux, G. Conforti, I. Gentil, and C. Léonard. Time reversal of diffusion processes under a finite entropy condition. *arXiv preprint arXiv:2104.07708*, 2021.

S. Chen, S. Chewi, J. Li, Y. Li, A. Salim, and A. R. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions, 2023.

T. Chen. On the importance of noise scheduling for diffusion models. *arXiv preprint arXiv:2301.10972*, 2023.

G. Conforti, A. Durmus, and M. G. Silveri. Score diffusion models without early stopping: finite fisher information is all you need, 2023.

References II

V. De Bortoli, J. Thornton, J. Heng, and A. Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021.

X. Gao, H. M. Nguyen, and L. Zhu. Wasserstein convergence guarantees for a general class of score-based generative models, 2023.

Q. Guo, S. Liu, Y. Yu, and P. Luo. Rethinking the noise schedule of diffusion-based generative models. *visible on Open Review*, 2023.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

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

T. Karras, M. Aittala, T. Aila, and S. Laine. Elucidating the design space of diffusion-based generative models, 2022.

H. Lee, J. Lu, and Y. Tan. Convergence for score-based generative modeling with polynomial complexity. *Advances in Neural Information Processing Systems*, 35:22870–22882, 2022.

References III

H. Lee, J. Lu, and Y. Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pages 946–985. PMLR, 2023.

A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8162–8171. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/nichol21a.html.

H. Robbins. An empirical bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 157–163. University of California Press, 1956.

J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

References IV

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations (ICLR)*, 2021.

P. Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. doi: 10.1162/NECO_a_00142.