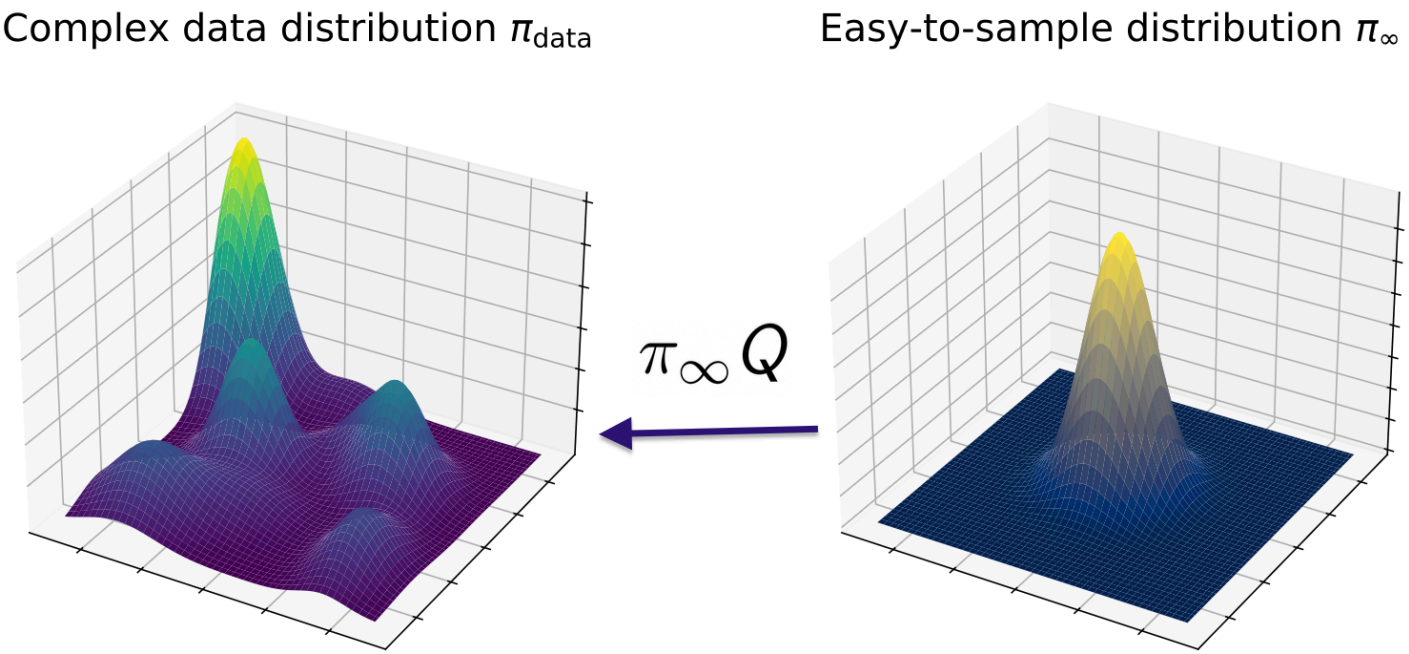
# Analysis of Noise Schedules in Score-Based Generative Modeling Algorithms

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

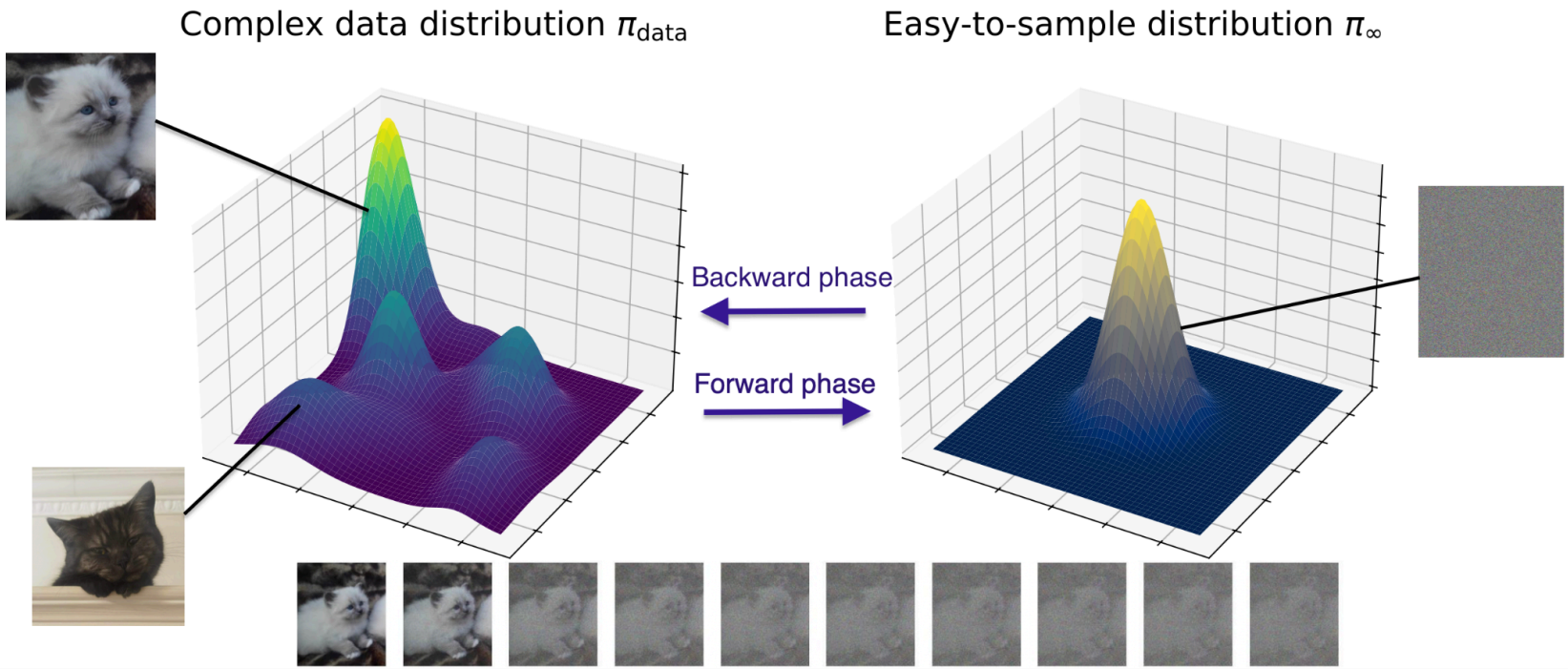
# Generative modelling framework

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



# SGMs Philosophy

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



# Introduction to SGMs: time reversal of diffusion processes

1. **Forward process**: Convert $\pi_{\text{data}}$ to an easy-to-sample distribution $\pi_\infty$ by progressively adding (Gaussian) noise.
   - Forward flow

$$\mathrm{d}\overrightarrow{X}_t = -\frac{1}{\sigma^2}\overrightarrow{X}_t \mathrm{d}t + \sqrt{2}\mathrm{d}B_t\,, \quad X_0 \sim \pi_{\text{data}}$$

   - Classical O.U. with time marginal, $Z \sim \mathcal{N}(0, \mathrm{I}_d)$, $Z \perp X_0$

$$\overrightarrow{X}_t = e^{-t}X_0 + \sqrt{1-e^{-2t}}Z\,.$$

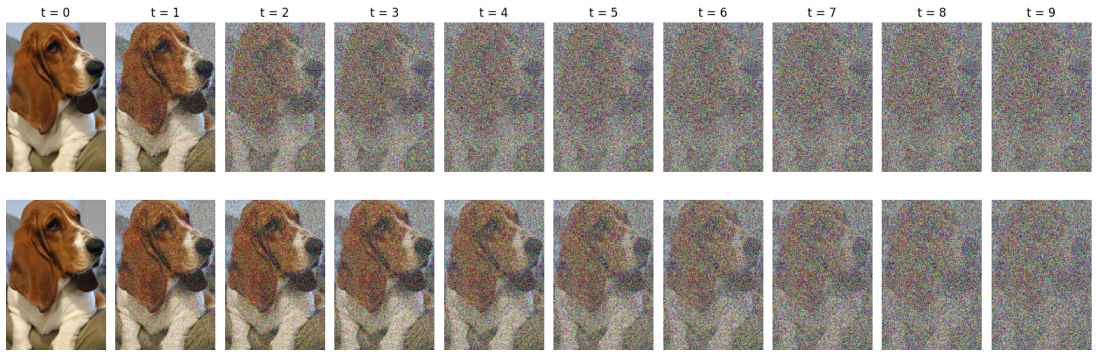
2. **Backward process**: Start from pure noise $\pi_\infty$ and reverse the noising dynamics to recover $\pi_{\text{data}}$. [Anderson, 1982]

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

$$\mathrm{d}\overleftarrow{X}_t = \left(\frac{1}{\sigma^2}\overleftarrow{X}_t + 2\underbrace{\nabla \log \pi_{T-t}\left(\overleftarrow{X}_t\right)}_{\textbf{score function}}\right)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t\,, \quad \overleftarrow{X}_0 \sim \pi_T.$$

# What is the appropriate amount of noise ?

- SGMs require to **hand-design** the intensity and the form of the progressive noising procedure.
- **Little is known theoretically** about the choice of a noise schedule. We only know **best practices** from experience.



# Adapted theoretical framework: time-inhomogeneous SDE

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**:

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

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

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

# SGMs in practice I: converging to $\pi_\infty$

- Forward process time-marginal writes as $X_t = m_t X_0 + \sigma_t Z$ with $m_t = \exp\{-\int_0^t \beta(s)\mathrm{d}s/(2\sigma^2)\}$ and $\sigma_t^2 = \sigma^2(1 - m_t^2)$.
- Recall that $\pi_{\mathrm{data}} = \pi_T Q_T$ but $\pi_t$ depends on $\pi_{\mathrm{data}}$,

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

- In practice we leverage the ergodicity of the O.U. kernel and set $T$ large so that with $\pi_\infty \sim \mathcal{N}\left(0, \sigma^2 \mathrm{I}_d\right)$,

$$\pi_{\mathrm{data}} \simeq \pi_\infty Q_T.$$

**Mixing time error.**

# SGMs in practice II: learn the score function

- The backward process depends on the score function $\nabla \log \pi_t(x)$ which is unknown.
- 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 $p_\tau(x)$ is **intractable !**
- Solution: its **conditional version** shares the same optimum

[Vincent, 2011]

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

**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**.
- We **discretize** $[0, T]$ by $N$ points with $t_k = kh$ with $h = T/N$, we let $t = t_k$ if $kh \leq t \leq (k+1)h$.
- Consider the **Exponential Integrator scheme**:

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

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

# KL upper bound with minimal hypothesis

## Theorem (S. et al 2024)

*Hyp: (i)* $\beta$ *is continuous, positive,* $\nearrow$*, with* $\int_0^\infty \beta(t)\mathrm{d}t = \infty$

*(ii) Novikov's condition on the difference of the actual and estimated scores.*

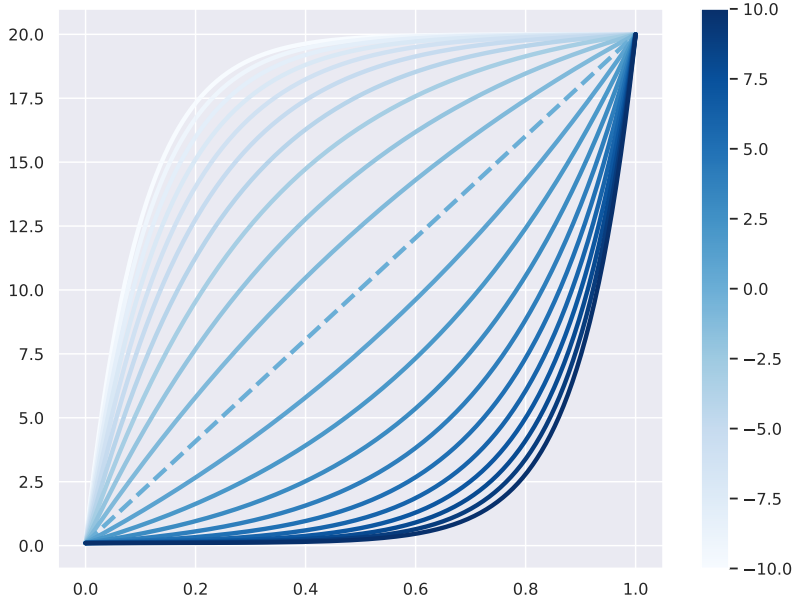
*(iii)* $\mathcal{I}(\pi_{\mathrm{data}}|\pi_\infty) < \infty$. *Then,*

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

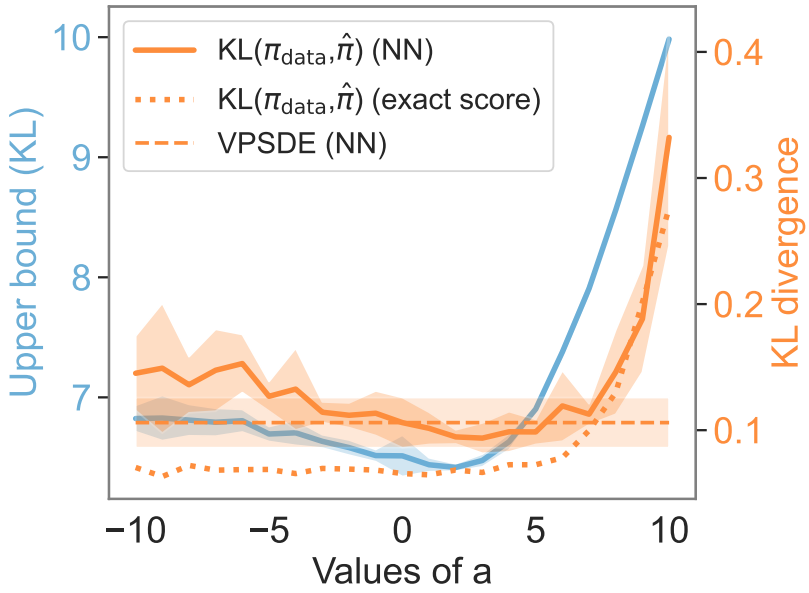
$$+\underbrace{\sum_{k=0}^{N-1}\mathcal{E}^{\beta}_{\theta,k}\int_{T-t_{k+1}}^{T-t_k}\beta(t)\mathrm{d}t}_{\textbf{Approx. error}} + \underbrace{2h\beta(T)\mathcal{I}(\pi_{\mathrm{data}}|\pi_\infty)}_{\textbf{Discr. error}}\,.$$

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

# Numerical analysis I



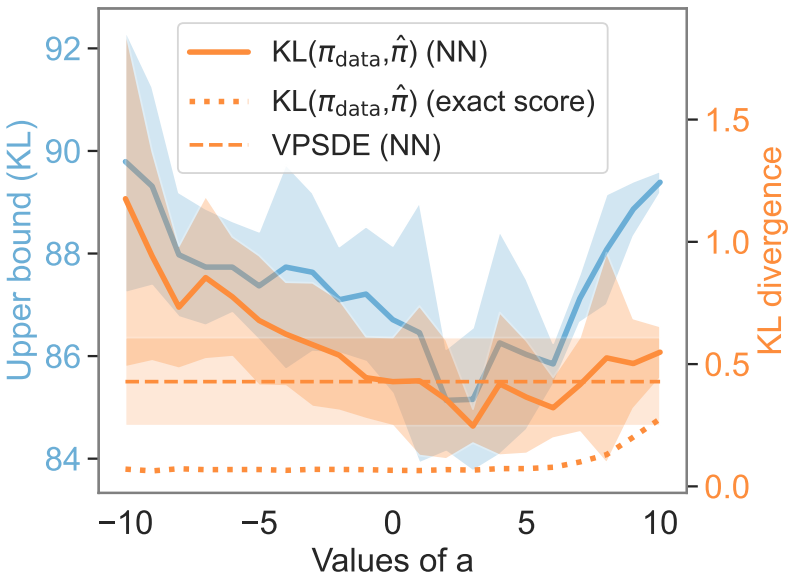

Noise schedules $\beta_a$



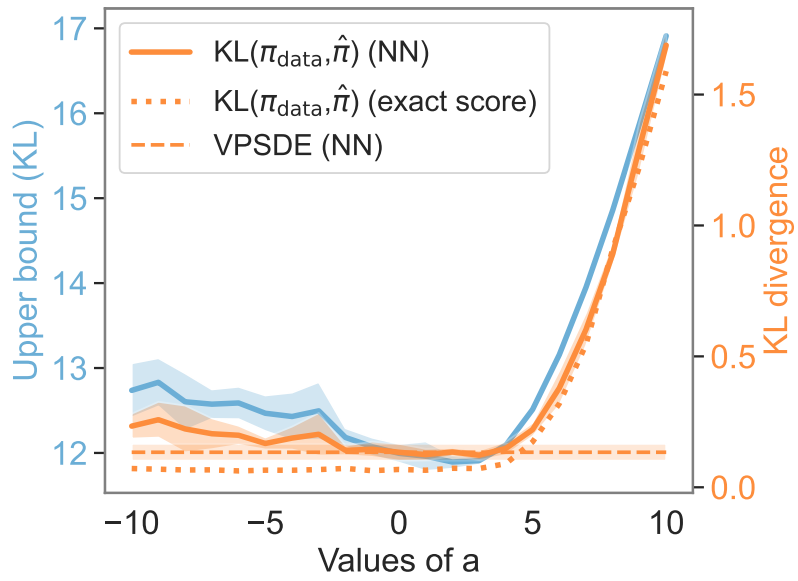

Isotropic Gaussian $\mathcal{N}(\mathbf{1}_d, 0.5\mathrm{I}_d)$

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

# Numerical analysis II




Anisotropic Gaussian




Correlated Gaussian

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

# Refined Wasserstein bound

## Theorem (S. et al 2024)

*Hyp:* *(i)* $\pi_t$ *is* $C_t$*-strongly log-concave through diffusion*
*(ii)* $\nabla \log \pi_t$ *is* $L_t$*-smooth through diffusion*
*(iii)* $\nabla \log \pi_t$ *is of linear growth of at most* $M$*.*

*Then,*

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

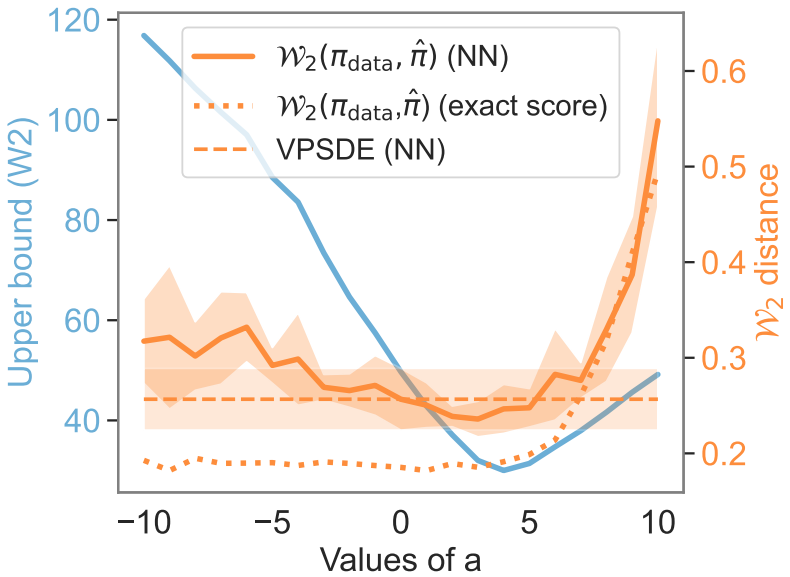
$$+ \sum_{k=0}^{N-1}\left(\int_{t_k}^{t_{k+1}} \bar{L}_t \bar{\beta}(t)\mathrm{d}t\right)\left(\frac{\sqrt{2h\beta(T)}}{\sigma} + \frac{h\beta(T)}{2\sigma^2} + \int_{t_k}^{t_{k+1}} 2\bar{L}_t\bar{\beta}(t)\mathrm{d}t\right) B$$

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

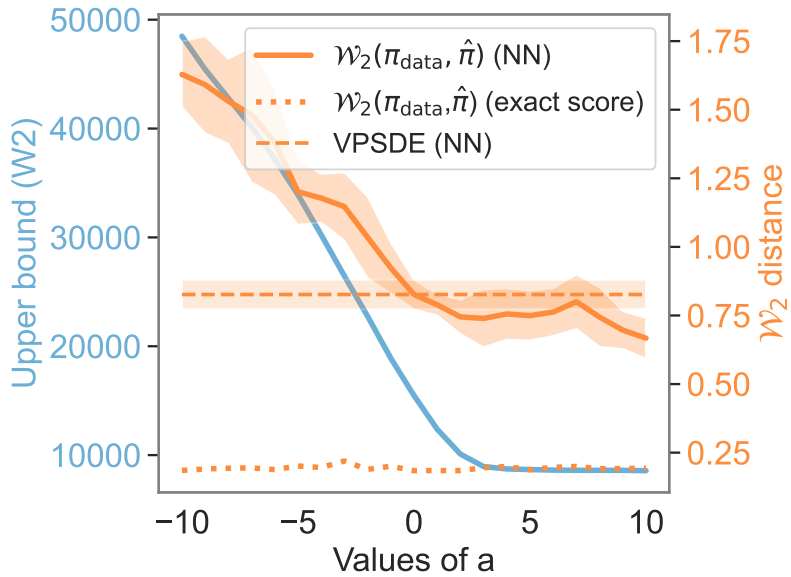
*with* $B = (\mathbb{E}[\|X_0\|^2] + \sigma^2 d)^{1/2}$ *and*

$$\mathcal{E}^\beta = \sup_{k\in\{0,..,N-1\}} \left\|\nabla \log \pi_{T-t_k}\left(\bar{X}^\theta_{t_k}\right) - s_\theta\left(T - t_k, \bar{X}^\theta_{t_k}\right)\right\|_{L_2}.$$

# Numerical analysis II




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



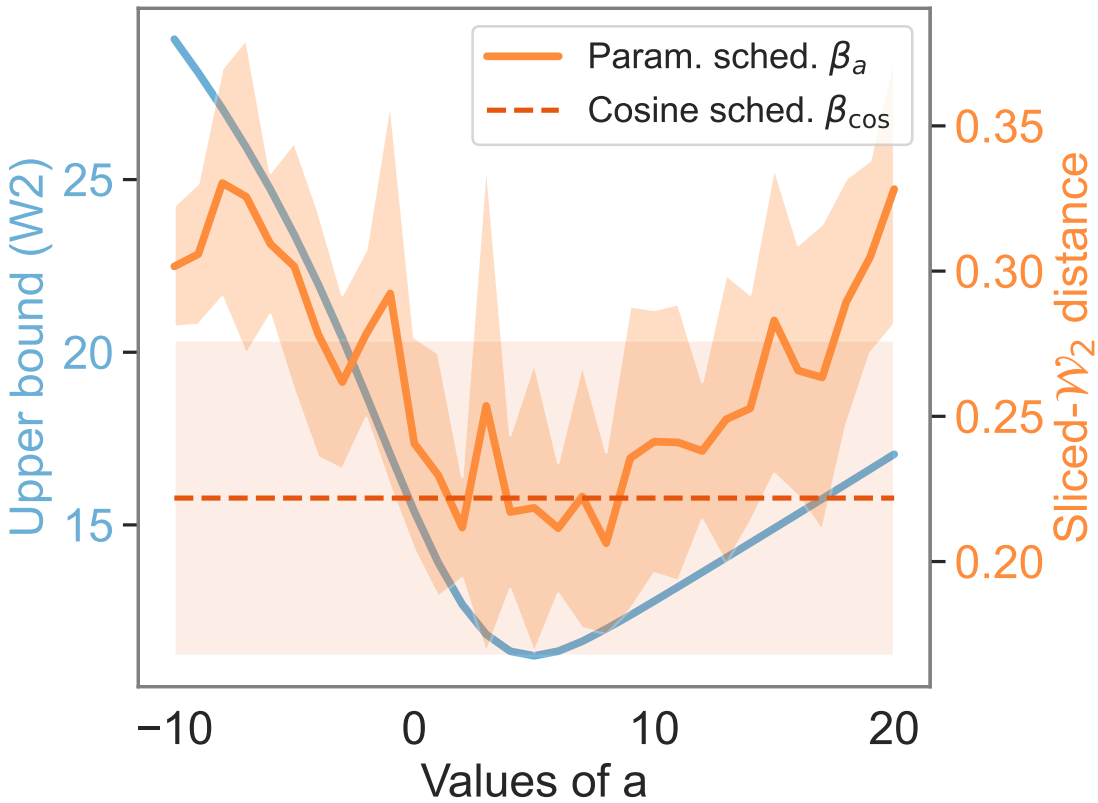

(b) Correlated setting $\mathcal{N}\left(\mathbf{1}_d, \Sigma^{(\text{corr})}\right)$ [1]

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

[1] $\Sigma^{(\text{corr})} \in \mathbb{R}^{d\times d}$ is a full matrix whose diagonal entries are equal to one and the off-diagonal terms are $\Sigma^{(\text{corr})}_{jj'} = 1/\sqrt{|j-j'|}$ for $1 \leq j \neq j' \leq d$

# Beyond the Gaussian setting

- Funnel distribution in dimension $d = 50$
  $\pi_{\text{data}}(x) = \mathcal{N}(x_1; 0, 1) \prod_{j=2}^{d} \mathcal{N}(x_j; 0, \exp(x_1))$




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