Most tImes Are uSed multiple tiMes (MIASM)

Sylvain Le Corff, Sorbonne Université, LPSM

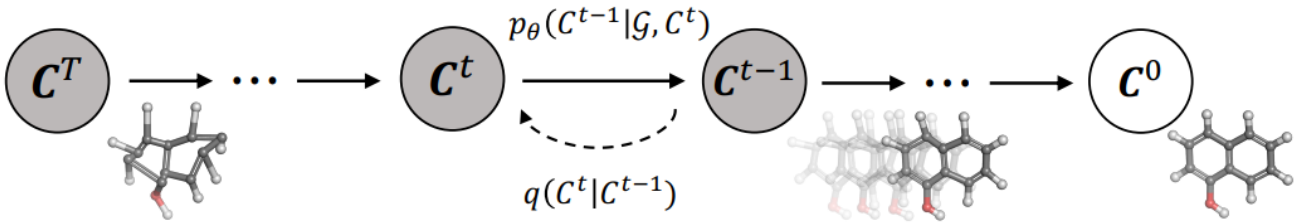
Based on joint works with
Stanislas Strasman, Claire Boyer, Vincent Lemaire (LPSM)
Antonio Ocello, Yazid Janati, Gabriel Cardoso, Eric Moulines (CMAP)

Generative modeling

Assumption: $(X^1, \ldots, X^N)$ in $\mathbb{R}^{d_x}$ are samples from some **unknown** distribution π_{data}.

What: generate synthetic instances of a target distribution π_{data}

Probabilistic model for generating molecular conformations (GeoDiff, Xu et al., 2022)

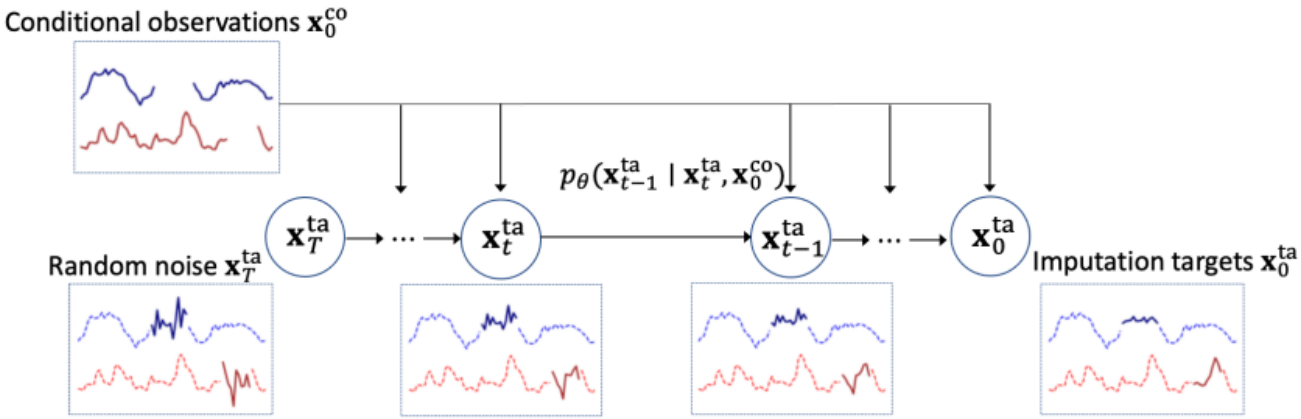


GEOM (37 million molecular conformations annotated by energy): generates new structures + chemical toolkit to calculate conformation energy.

MOst times are USed multiple timEs (MOUSE)

Generative modeling

Probabilistic time series imputation (CSDI, Tashiro et al., 2021)



Healthcare dataset in PhysioNet Challenge 2012 (4000 clinical time series with 35 variables for 48 hours from intensive care unit).

Synthetic Data Generation for Privacy and Security (TabDDPM, Kotelnikov et al., 2023), etc.

most Times ARe useD multiple times (TARDigrade)

Generative modeling

PierrE pictURes (PEUR) dataset (3 pictures, dataset being built slowly but surely).

Improving corrupted data

Generative modeling

PierrE pictURes (PEUR) dataset (3 pictures, dataset being built slowly but surely).

Improving corrupted data

Generative modeling

PierrE pictURes (PEUR) dataset (3 pictures, dataset being built slowly but surely).

Improving corrupted data

most times are used multiple times (juste pour le plaisir)

Generative modeling

(1) Estimate π_{data} with a **parametric** probability distribution p_θ.

1. Choose a suitable parametric form for p_θ.
   $\rightsquigarrow$ p_θ is parameterized using a Neural Network.
   $\rightsquigarrow$ $\mathrm{p}_\theta \geq 0$, $\int \mathrm{p}_\theta = 1 \rightarrow$ **constraints** on the modelization.

2. Train p_θ to approximate π_{data} using the samples $(X^1, \dots, X^N) \sim \pi_{\text{data}}$: $\mathcal{L}(\theta) = \sum_{i=1}^{N} -\log \mathrm{p}_\theta(X^i)$.
   $\rightsquigarrow$ Minimize $\mathcal{L}(\theta)$ or an upperbound $\rightarrow$ find optimal parameter θ_*.

mOst times Used aRe multiple timeS (OURS)

Controlled generation

(2) Perform controlled generation using p_{θ_*}.

$\rightsquigarrow$ Target distribution: weight p_{θ_*} with a function $x \mapsto g(x)$

$$\phi(\mathrm{d}x) = \frac{g(x)\mathsf{p}_{\theta_*}(\mathrm{d}x)}{\int g(z)\mathsf{p}_{\theta_*}(\mathrm{d}z)},$$

$\rightsquigarrow$ Posterior sampling: $g(x) = p(y|x)$.

mosT timEs aRe used MultIple TimES (TERMITES)

Score-Based Generative Models (SGMs)

What: generate synthetic instances of a target distribution π_{data}

Why: challenges in modeling the complexity of real data, preventing conventional parametric modeling or traditional maximum likelihood methods.

> *Creating noise from data is easy; creating data from noise is generative modeling.* (Song et al., *Score-Based Generative Modeling through Stochastic Differential Equations*)

Who: SGMs address this by

1. (forward phase) introducing progressively noise into the samples,
2. (backward phase) reversing the noising dynamics, with the help of a score function usually learned using deep neural networks.

mOst times aRe usEd multiple times (OREortyx)

Noise scheduling tuning

Practitioners' corner: time-inhomogeneous SGMs - the noise schedule has central role, as exhibited in numerical experiments.

[Chen et al., 2023]: performance of SGMs relies on the chosen noise scheduling and optimal strategy varies depending on the task, $e.g.$, image sizes.

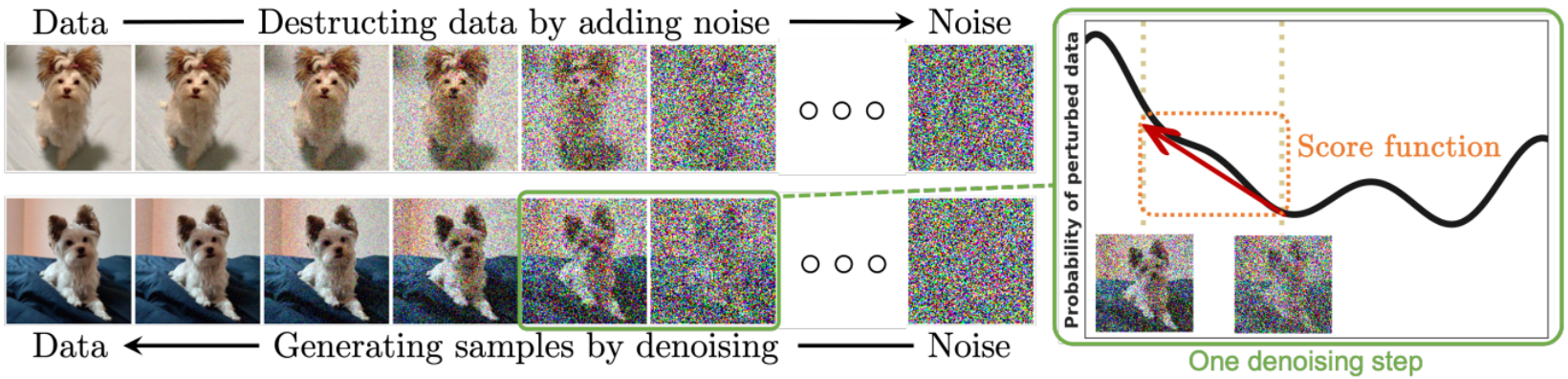


Figure: From Yang et al., 2023

MOst times are used multiple Times (MOTh)

Estimating π_{data} with diffusion models

Diffusion models: the forward process - DDPM [Ho et al., 2020]

Consider the *forward noising* process

$$X_k = \sqrt{1-\beta_k} X_{k-1} + \sqrt{\beta_k} Z_k\,, \quad \beta_k \in [0,1], \quad X_0 \sim \pi_{\text{data}}\,,$$

where $Z_k \sim \mathcal{N}(\mathbf{0}_{d_x}, \mathrm{I}_{d_x})$.

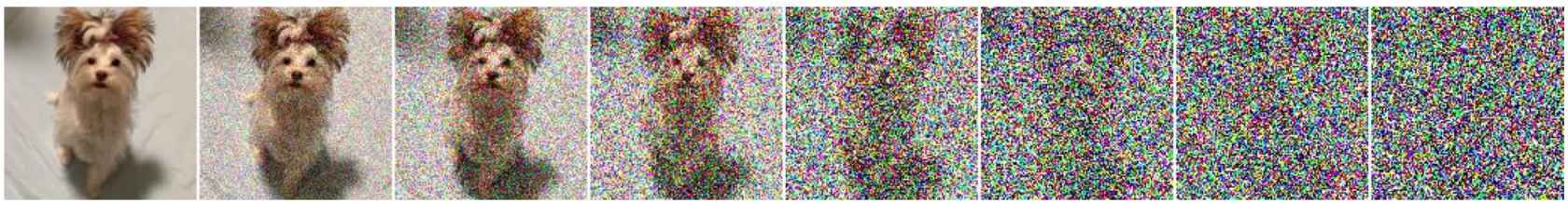


Figure: One sample $X_{0:n}$.

$X_k \sim \pi_k$ where [1] $\pi_k(\mathrm{d}x_k) := \int \pi_{\text{data}}(\mathrm{d}x_0)\mathcal{N}(\mathrm{d}x_k; \sqrt{\bar{\alpha}_k}x_0, (1-\bar{\alpha}_k)\mathrm{I}_{d_x})$.

$(X_k)_{k\geq 0}$ is a discrete-time OU process.

MOst times are used muLtiple times (MOLa)

[1] $\bar{\alpha}_k := \prod_{j=1}^{k}(1-\beta_j)$.

Diffusion models: the backward process

Note that $\pi_{1:n|0}(x_{1:n}|x_0) = \pi_{n|0}(x_n|x_0) \prod_{k=2}^{n} \pi_{k-1|0,k}(x_{k-1}|x_0, x_k)$, where $\pi_{n|0}(x_n|x_0) = \mathcal{N}(x_n; \bar{\alpha}_n^{1/2} x_0, (1 - \bar{\alpha}_n)\mathrm{I})$ and

$$\pi_{k-1|0,k}(x_{k-1}|x_0, x_k) = \mathcal{N}\left(x_{k-1}; \boldsymbol{\mu}_k(x_0, x_k), \sigma_k^2 \mathrm{I}_d\right) ,$$

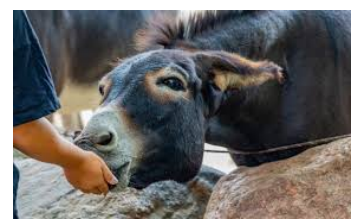
with

$$\boldsymbol{\mu}_k(x_0, x_k) = \bar{\alpha}_{k-1}^{1/2} x_0 + (1 - \bar{\alpha}_{k-1} - \sigma_k^2)^{1/2} (x_k - \bar{\alpha}_k^{1/2} x_0) / (1 - \bar{\alpha}_k)^{1/2}.$$

$\rightsquigarrow$ We know how to write the joint distribution of $X_{1:n}$ given X_0.

$\rightsquigarrow$ Use this decomposition to turn noise into samples from π_0.

$$\mathsf{p}_{0:n}^{\theta}(\mathrm{d}x_{0:n}) = \mathsf{p}_n(\mathrm{d}x_n) \prod_{k=0}^{n-1} p_k^{\theta}(\mathrm{d}x_k|x_{k+1}) .$$



Most times are Used muLtiple timEs (MULE)

Diffusion models: the backward process

$\rightsquigarrow$ Use this decomposition to turn noise into samples from π_0.

$$\mathsf{p}^{\theta}_{0:n}(\mathrm{d}x_{0:n}) = \mathsf{p}_n(\mathrm{d}x_n) \prod_{k=0}^{n-1} p^{\theta}_k(\mathrm{d}x_k|x_{k+1}) \,,$$

where p_n is a std Gaussian and

$$p^{\theta}_k(\mathrm{d}x_k|x_{k+1}) = \mathcal{N}(\mathrm{d}x_k; \mu^{\theta}_{k+1}(x_{k+1}), \beta_{k+1}\mathrm{I}_{d_x})$$

with $\mu^{\theta}_{k+1}(x_{k+1})$ obtained by replacing x_0 in $\boldsymbol{\mu}_{k+1}(x_0, x_{k+1})$ with a prediction

$$\hat{x}_{0|k,\theta}(x_{k+1}) := \bar{\alpha}_{k+1}^{-1/2} \left(x_{k+1} - (1-\bar{\alpha}_{k+1})^{1/2} \mathbf{e}^{\theta}(x_{k+1}, k+1) \right) .$$

Most times are Used multiple timeS (MUSkox)

Diffusion models: the backward process

We use

$$p_k^\theta(\mathrm{d}x_k|x_{k+1}) = \mathcal{N}(\mathrm{d}x_k; \mu_{k+1}^\theta(x_{k+1}), \beta_{k+1}\mathrm{I}_{d_x})$$

with $\mu_{k+1}^\theta(x_{k+1})$ obtained by replacing x_0 in $\boldsymbol{\mu}_{k+1}(x_0, x_{k+1})$ with a prediction

$$\hat{x}_{0|k,\theta}(x_{k+1}) := \bar{\alpha}_{k+1}^{-1/2}\left(x_{k+1} - (1-\bar{\alpha}_{k+1})^{1/2}\mathbf{e}^\theta(x_{k+1}, k+1)\right) .$$

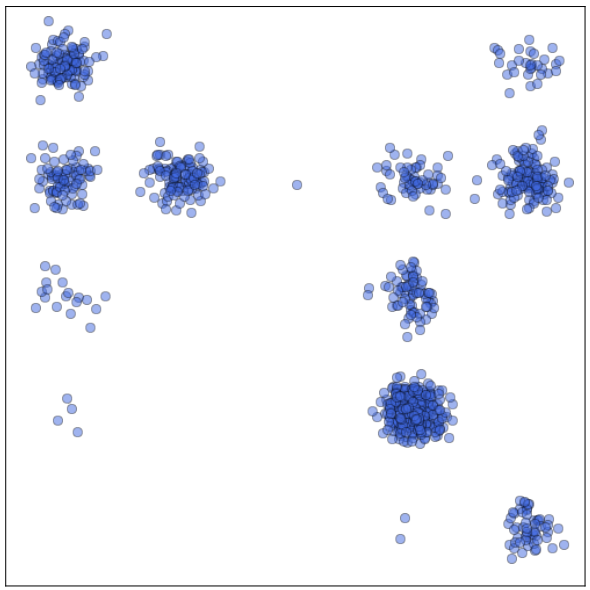
$\mathbf{e}^{\theta *}(X_t, t)$ might be seen as the predictor of the noise added to X_0 to obtain X_t (in the forward pass) and justifies the *prediction* terminology.

The parameter θ is obtained by minimizing a variational loss between the forward and backward joint distributions.

moSt timEs Are used muLtiple times (SEA Lion)

Diffusion models: an illustration

Data distribution π_{data}



Marginal p_{500}

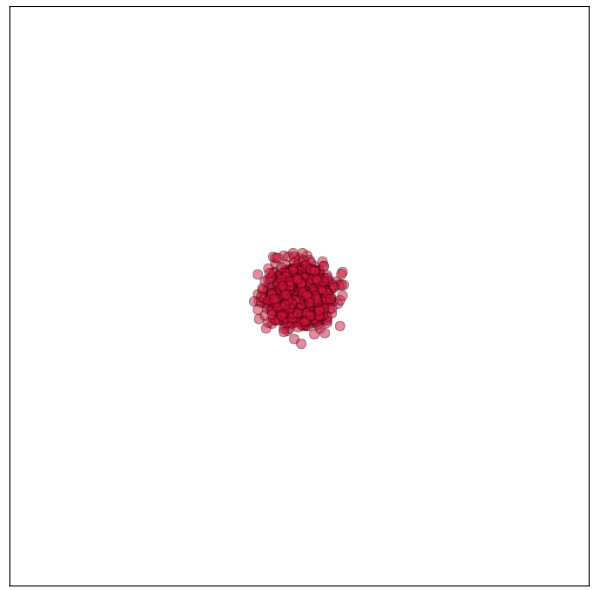
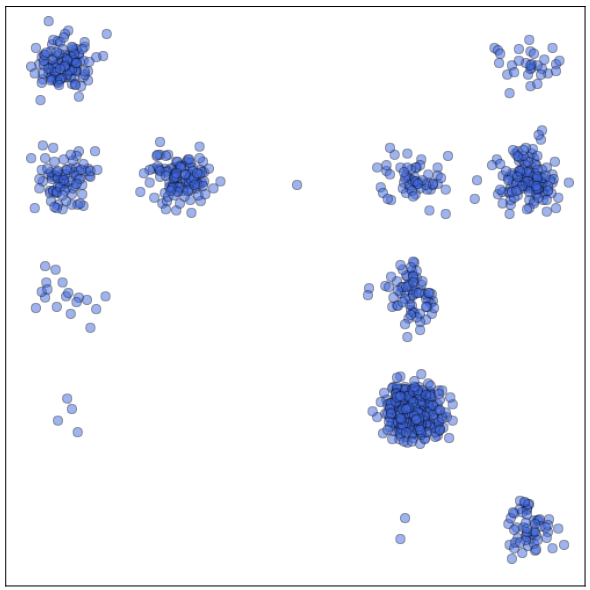


Figure: Samples from p_t for some time steps ranging from $n := 500$ to 1. π_0 is a Gaussian mixture.

Diffusion models: an illustration

Data distribution π_{data}



Marginal p_{150}

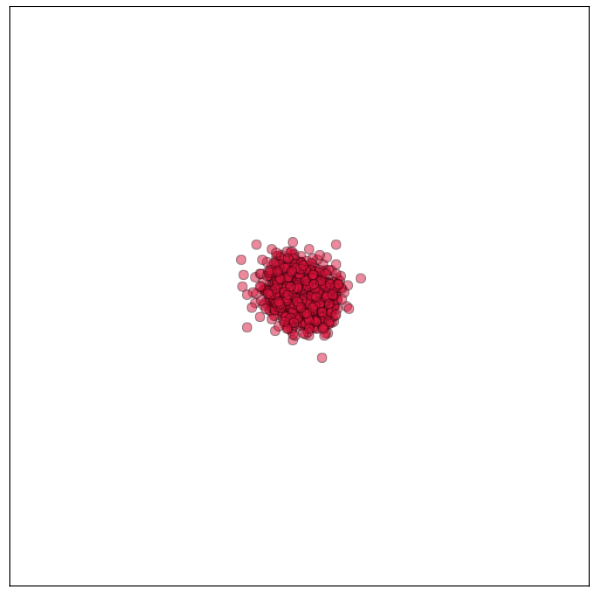
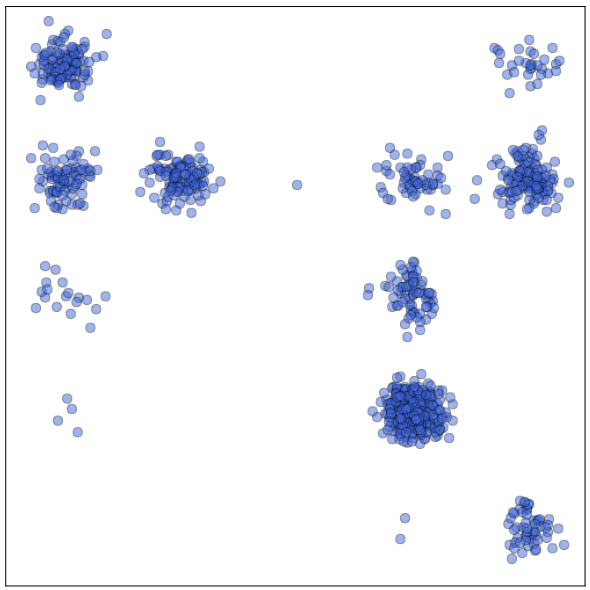


Figure: Samples from p_t for some time steps ranging from $n := 500$ to 1. π_0 is a Gaussian mixture.

Diffusion models: an illustration

Data distribution π_{data}



Marginal p_{100}

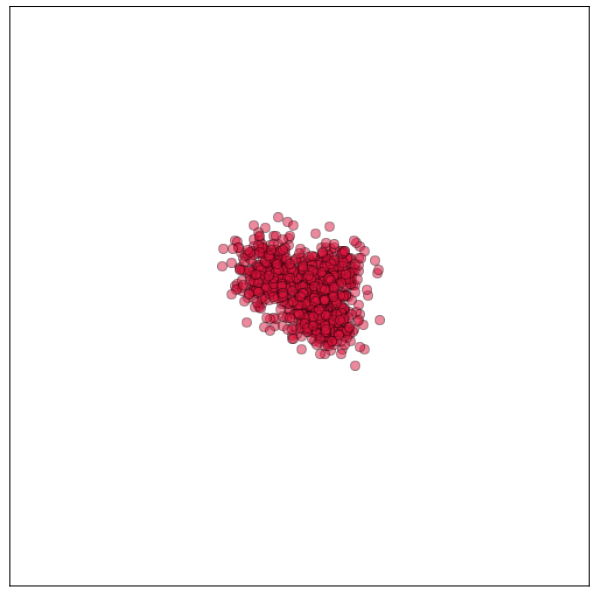
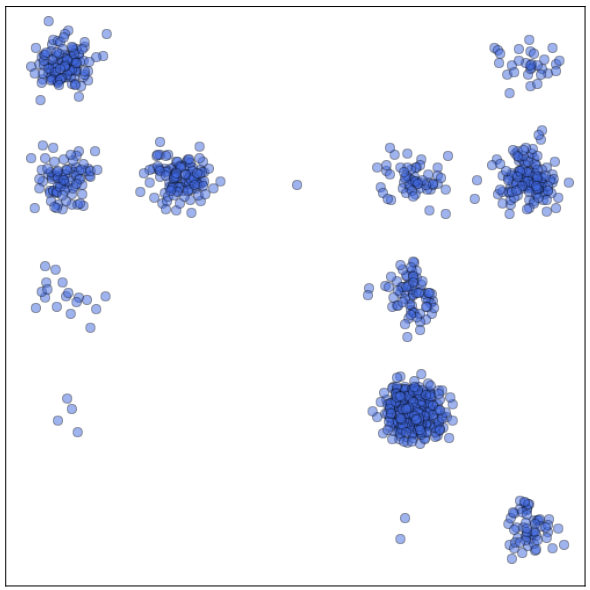


Figure: Samples from p_t for some time steps ranging from $n := 500$ to 1. π_0 is a Gaussian mixture.

Diffusion models: an illustration

Data distribution π_{data}



Marginal p_{80}

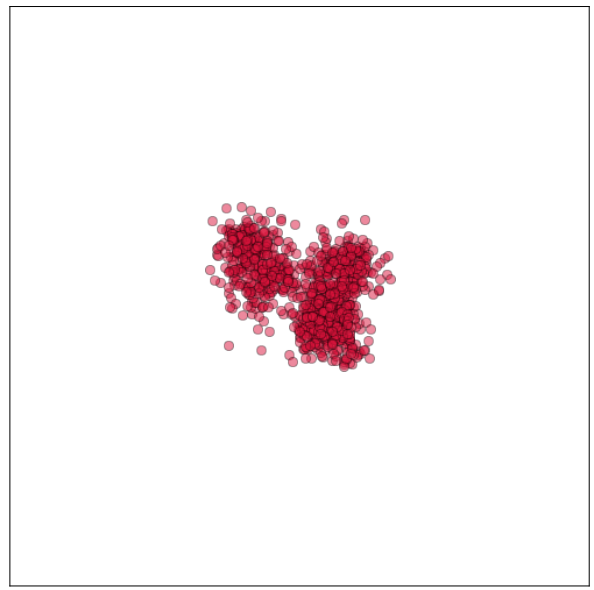
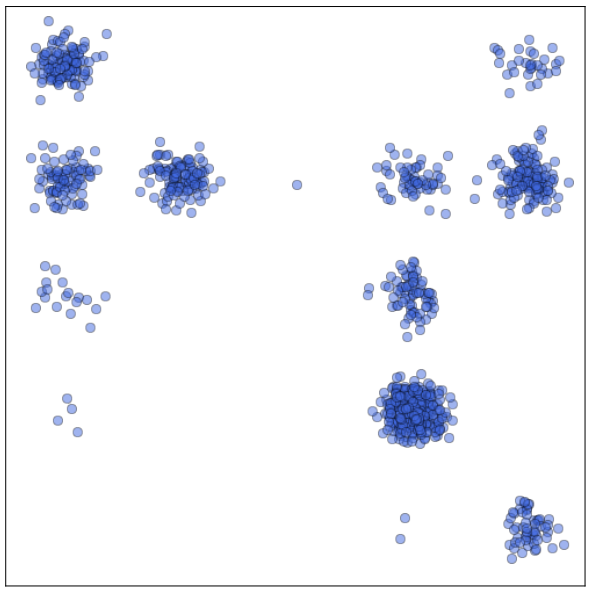


Figure: Samples from p_t for some time steps ranging from $n := 500$ to 1. π_0 is a Gaussian mixture.

moSt timEs Are used multiple timeS (SEA Slug)

Diffusion models: an illustration

Data distribution π_{data}



Marginal p_{70}

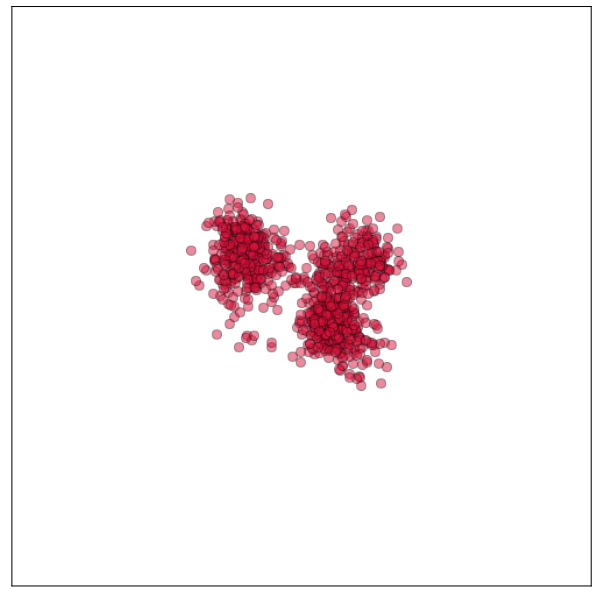
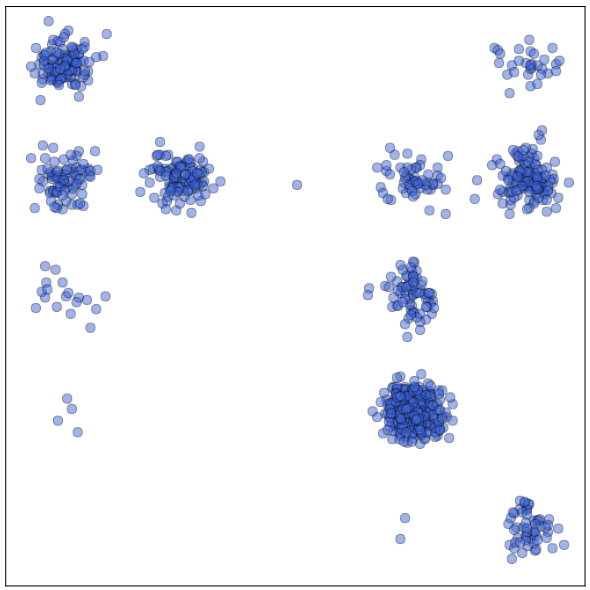


Figure: Samples from p_t for some time steps ranging from $n := 500$ to 1. π_0 is a Gaussian mixture.

Diffusion models: an illustration

Data distribution π_{data}



Marginal p_{50}

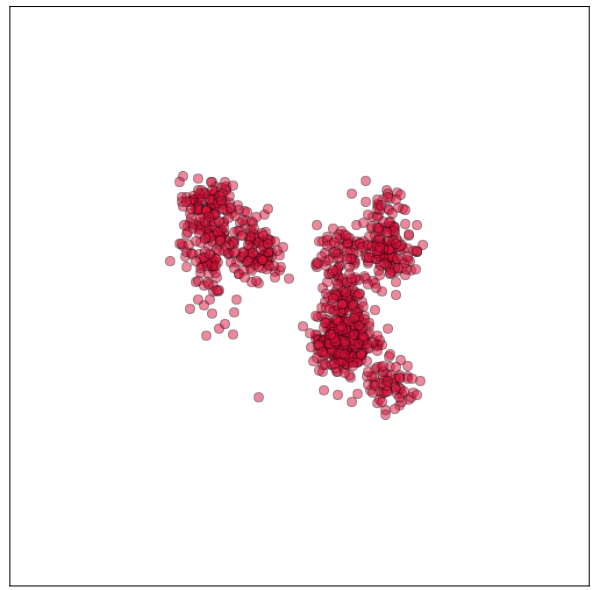
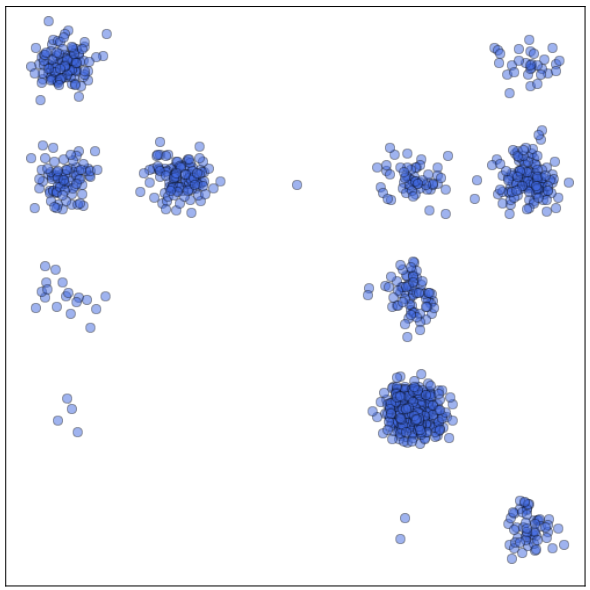


Figure: Samples from p_t for some time steps ranging from $n := 500$ to 1. π_0 is a Gaussian mixture.

Diffusion models: an illustration

Data distribution π_{data}



Marginal p_{40}

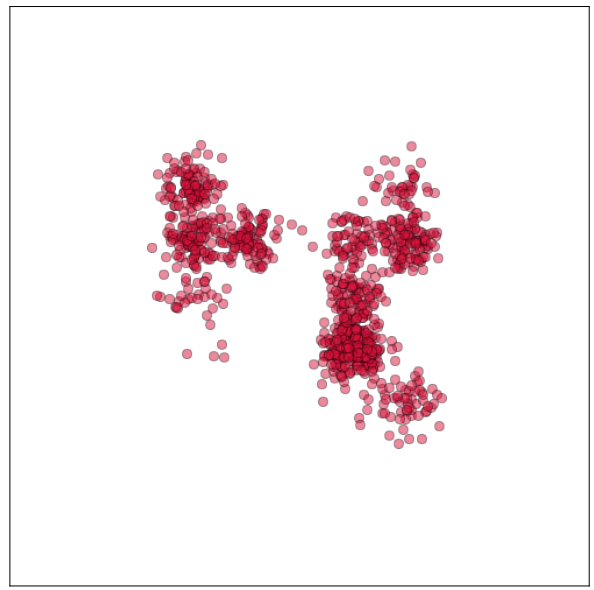
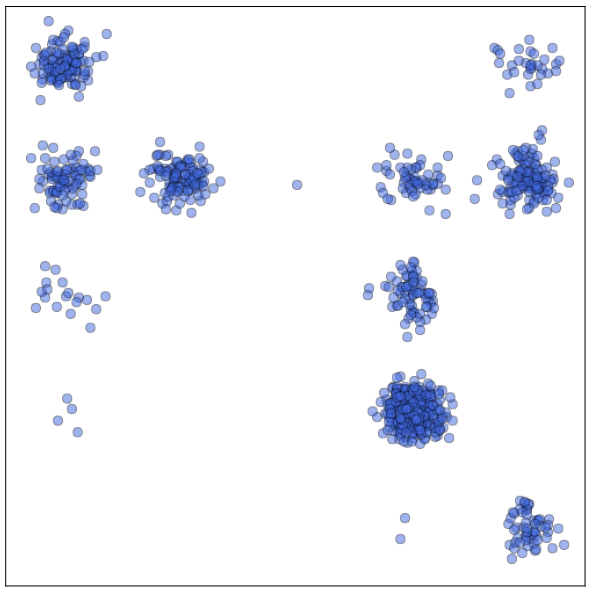


Figure: Samples from p_t for some time steps ranging from $n := 500$ to 1. π_0 is a Gaussian mixture.

Diffusion models: an illustration

Data distribution π_{data}



Marginal p_{20}

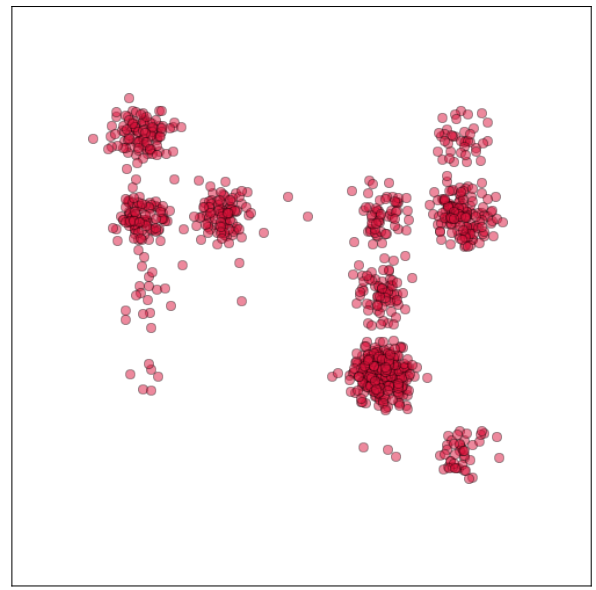
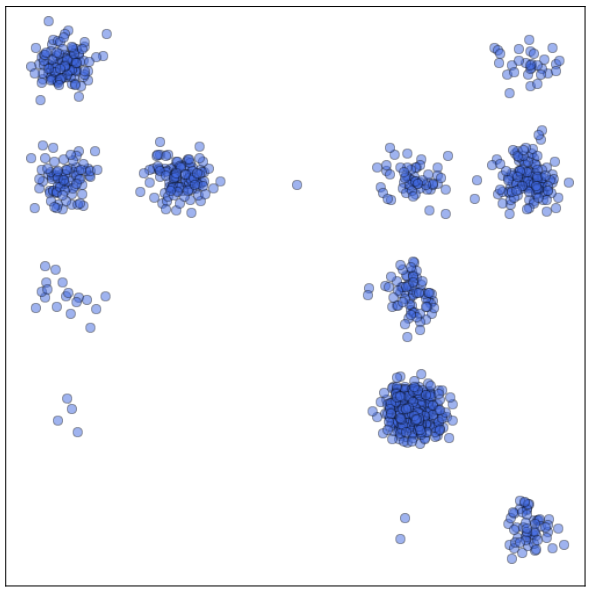


Figure: Samples from p_t for some time steps ranging from $n := 500$ to 1. π_0 is a Gaussian mixture.

moSt timEs Are used multiple timeS (SEA Slug)

Diffusion models: an illustration

Data distribution π_{data}



Marginal p_{15}

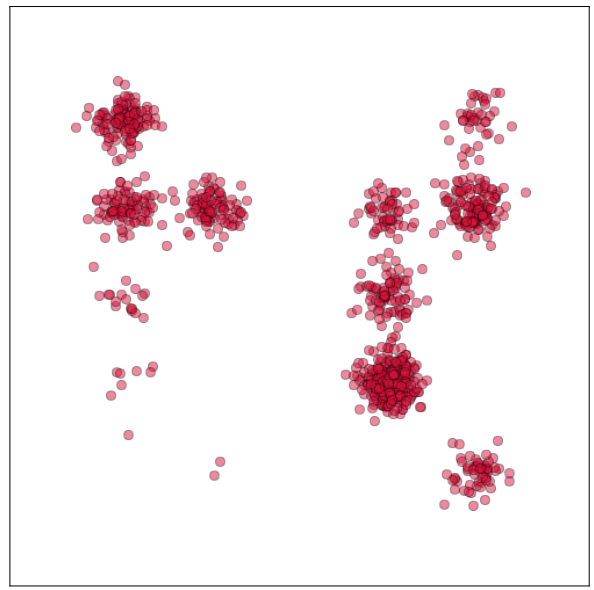
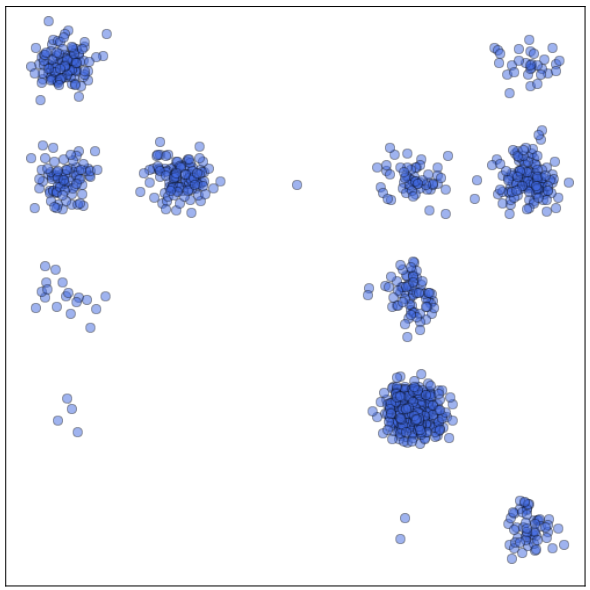


Figure: Samples from p_t for some time steps ranging from $n := 500$ to 1. π_0 is a Gaussian mixture.

Diffusion models: an illustration

Data distribution π_{data}



Marginal p_5

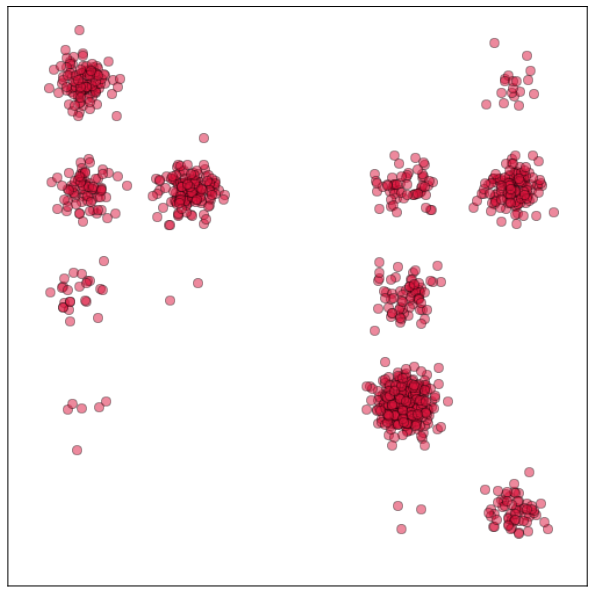


Figure: Samples from p_t for some time steps ranging from $n := 500$ to 1. π_0 is a Gaussian mixture.

Diffusion models: convergence

In the first theoretical results: strong assumptions on the data distribution (polynomial growth of the score), (de Bortoli et al., 2021):

$$\left\|\pi_{\mathrm{data}} - p_0^\theta\right\|_{\mathrm{tv}} \leq c_0 M \exp(c_1 n) + c_2 \left(n^{-1} + n^{-1/2}\right) ,$$

where M **quantifies the quality of the score approximation**.

In most recent works, we only require π_{data} to have a finite relative Fisher information w.r.t the standard Gaussian distribution, (Conforti et al., 2023):

$$\mathrm{KL}(\pi_{\mathrm{data}}, p_0^\theta) \leq \exp(-c_0 n)\mathrm{KL}(\mathcal{N}(0, I_d), \pi_0) + Mn + c_1 h \,,$$

Assuming that $\mathbb{E}_{\pi_{\mathrm{data}}}[\|\nabla \log(\mathrm{d}\pi_{\mathrm{data}}/\mathrm{d}\gamma_d)\|^2] < \infty$.

moSt times Are used multiple TimEs (SATanic leaf-tailed gEcko)

Guarantees on the approximation of π_{data}
Score-based training procedures

SGM - Forward phase

Data noised using the Ornstein–Uhlenbeck (OU) process:

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

Fix $T > 0$, then,

$$\mathrm{KL}\left(\mathcal{L}\left(\overrightarrow{X}_T\right), \pi_\infty\right) \lesssim \exp\left(-\frac{1}{2\sigma^2}\int_0^T \beta(s)\mathrm{d}s\right)\mathrm{KL}\left(\pi_{\mathsf{data}}, \pi_\infty\right)$$

with

$$\mathrm{KL}\left(\mu, \nu\right) := \int \log\left(\frac{\mathrm{d}\mu(x)}{\mathrm{d}\nu(x)}\right)\mu(\mathrm{d}x)\,.$$

Fokker-Planck for $(p_t)_{0\leq t\leq T}$ + logarithmic Sobolev inequality + Gronwall's inequality

moSt tImeS Are uSed multiple Times (ScISSor-tailed flycaTcher)

Time reversal

For $p_t := \mathcal{L}(\overrightarrow{X}_t)$, $(t,x) \mapsto \nabla \log p_t(x)$ is the **score function**. We consider the time reversal of the forward process, $i.e.$, the process satisfying

$$\mathrm{d}\overleftarrow{X}_t = \left(\frac{\bar{\beta}(t)}{2\sigma^2} \overleftarrow{X}_t + \bar{\beta}(t) \nabla \log p_{T-t} \left(\overleftarrow{X}_t \right) \right) \mathrm{d}t + \sqrt{\bar{\beta}(t)} \mathrm{d}B_t, \quad \overleftarrow{X}_0 \sim p_T \,,$$

with $\bar{\beta}(t) := \beta(T-t)$. It satisfies

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

In our setting,

$$\mathcal{L}\left(\overrightarrow{X}_T \right) = \mathcal{L}\left(\overleftarrow{X}_0 \right) \approx \pi_\infty \,, \qquad \mathcal{L}\left(\overleftarrow{X}_T \right) = \mathcal{L}\left(\overrightarrow{X}_0 \right) \approx \pi_{\text{data}} \,.$$

moSt timEs Are uSed multiPle tImes (SEA SPIder)

Modified score

Let $\tilde{p}_t := p_t/\varphi_{\sigma^2}$, with φ_{σ^2} the density of π_∞. With respect to the **modified score function**, the backward dynamics is

$$\mathrm{d}\overleftarrow{X}_t = \left(-\frac{\bar{\beta}(t)}{2\sigma^2}\overleftarrow{X}_t + \bar{\beta}(t)\nabla \log \tilde{p}_{T-t}\left(\overleftarrow{X}_t\right)\right)\mathrm{d}t + \sqrt{\bar{\beta}(t)}\mathrm{d}B_t, \quad \overleftarrow{X}_0 \sim p_T\,.$$

If the score (or the modified score) is known, we can (in theory) simulate the backward process and get *data from noise*.

Let $s_\theta : [0,T] \times \mathbb{R}^d \mapsto \mathbb{R}^d$ be such that

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

with $\tau \sim \mathcal{U}(0,T)$ independent of the forward process $(\overrightarrow{X}_t)_{t\geq 0}$.

mosT times aRE usEd multiple times (TREE creeper)

Discretization scheme

As the linear part can be simulated exactly, we consider the Exponential Integrator scheme:

Let $0 =: t_0 \leq t_1 \leq \cdots \leq t_N := T$. Consider

$$\mathrm{d}\overleftarrow{X}^\theta_t = \bar{\beta}(t)\left(-\frac{1}{2\sigma^2}\overleftarrow{X}^\theta_t + \tilde{s}_\theta\left(T - t_k, \overleftarrow{X}^\theta_{t_k}\right)\right)\mathrm{d}t + \sqrt{\bar{\beta}(t)}\mathrm{d}B_t\,,$$

for $t \in [t_k, t_{k+1})$, with $\tilde{s}_\theta(t,x) := s_\theta(t,x) + x/\sigma^2$, and $\overleftarrow{X}^\theta_0 \sim \pi_\infty$.

We denote $\widehat{\pi}_N^{(\beta,\theta)}$ the marginal probability density of $\overleftarrow{X}^\theta_T$.

The loss function is built using the conditional score:

$$\mathcal{L}(\theta) = \mathbb{E}\left[\alpha_\tau \left\|\nabla \log p_{\tau|0}(X_\tau|X_0) - s_\theta(\tau, X_\tau)\right\|^2\right]\,.$$

most times aRe used multiple tImes (RIver otter)

Assumptions

H1 The noise schedule is continuous, non decreasing and such that

$$\int_0^\infty \beta(t)\mathrm{d}t = \infty\,.$$

H2 The data distribution has finite Fisher information w.r.t. the normal distribution, $i.e.$,

$$\mathcal{I}(\pi_{\mathrm{data}}|\pi_\infty) := \int \left\| \nabla \log \left(\frac{\mathrm{d}\pi_{\mathrm{data}}}{\mathrm{d}\pi_\infty} \right) \right\|^2 \mathrm{d}\pi_{\mathrm{data}} < \infty\,.$$

H3 The parameter θ and the schedule β satisfy

$$\mathbb{E}\left[\exp\left\{ \frac{1}{2} \int_0^T \bar{\beta}(t) \left\| \left(\tilde{s}\left(T-t, \overleftarrow{X}_t\right) - \tilde{s}_\theta\left(T-t_k, \overleftarrow{X}_{t_k}\right)\right)\right\|^2 \mathrm{d}t \right\}\right] < \infty\,.$$

most times aRe used multiple tImes (RIver otter)

Why it works

$\bar{\mathbb{Q}}_N^{\beta,\theta} \in \mathcal{P}(C([0,T],\mathbb{R}^d))$: path measure (backward diffusion). $\widehat{\pi}_N^{(\beta,\theta)}$ the marginal probability density of $\overleftarrow{X}_T^\theta$.

By the data processing inequality,

$$\mathrm{KL}\left(\pi_{\mathrm{data}} \middle|\middle| \widehat{\pi}_N^{(\beta,\theta)}\right) = \mathrm{KL}\left(p_T Q_T \middle|\middle| \widehat{\pi}_N^{(\beta,\theta)}\right) \leq \mathrm{KL}\left(p_T \mathbb{Q}_T \middle|\middle| \pi_\infty \bar{\mathbb{Q}}_N^{\beta,\theta}\right) .$$

By applying Girsanov theorem,

$$\mathrm{KL}\left(\pi_{\mathrm{data}} \| \widehat{\pi}_N^{(\beta,\theta)}\right) \leq \mathrm{KL}\left(p_T \| \varphi_{\sigma^2}\right) + \frac{1}{2}\int_0^T \bar{\beta}(t)\mathbb{E}\left[\left\| \nabla \log \tilde{p}_{\tau_t}\left(\overleftarrow{X}_t\right) - \tilde{s}_\theta\left(\tau_k, \overleftarrow{X}_{t_k}\right)\right\|^2\right] \mathrm{d}t .$$

most times aRe used multiple tImes (RIver otter)

Upper bound

Theorem

Assume that H1, H2 and H3 hold. Then,

$$\mathrm{KL}\left(\pi_{\mathrm{data}} \middle|\middle| \widehat{\pi}_N^{(\beta,\theta)}\right) \leq \mathcal{E}_1(\beta) + \mathcal{E}_2(\theta,\beta) + \mathcal{E}_3(\beta)\,,$$

where

$$\begin{aligned}
\mathcal{E}_1(\beta) &:= \mathrm{KL}\left(\pi_{\mathrm{data}} || \pi_\infty\right) \exp\left\{-\frac{1}{\sigma^2}\int_0^T \beta(s)\mathrm{d}s\right\}\,,\\
\mathcal{E}_2(\theta,\beta) &:= \sum_{k=1}^{N}\int_{t_k}^{t_{k+1}} \beta(t)\mathrm{d}t\mathbb{E}\left[\left\|\nabla \log \tilde{p}_{T-t_k}\left(\overrightarrow{X}_{T-t_k}\right) - \tilde{s}_\theta\left(T-t_k, \overrightarrow{X}_{T-t_k}\right)\right\|^2\right]\,,\\
\mathcal{E}_3(\beta) &:= 2h\beta(T)\,\mathcal{I}(\pi_{\mathrm{data}}|\pi_\infty)\,,
\end{aligned}$$

with $h := \sup_{k\in\{1,\ldots,N\}}(t_k - t_{k-1})$ *and* $t_0 := 0$.

most times aRe used multiple tImes (RIver otter)

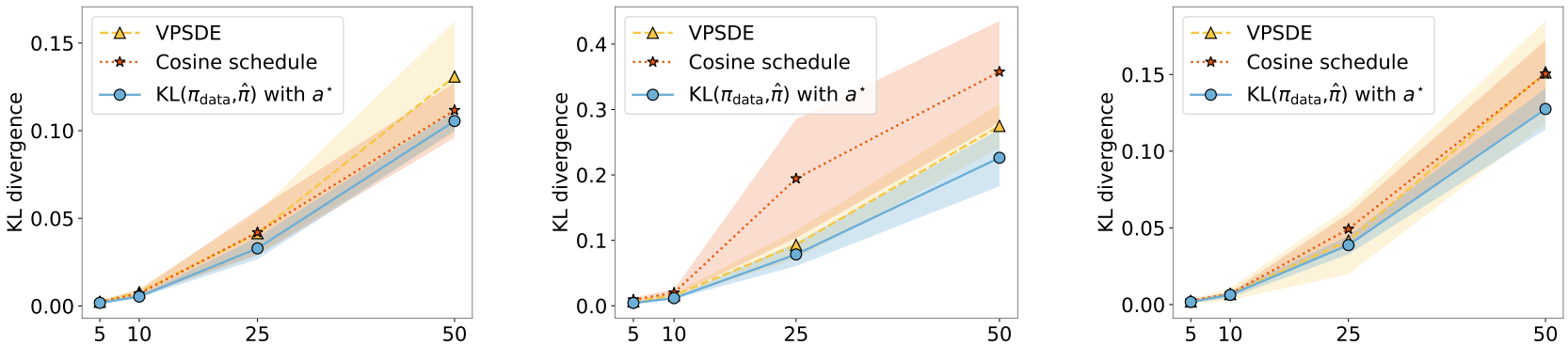
Gaussian case

Let the true distribution be Gaussian in dimension $d = 50$ with mean $\mathbf{1}_d$ and different choices of covariance structure.

1. (Isotropic) $\Sigma^{(\mathrm{iso})} = 0.5\mathrm{I}_d$.
2. (Heteroscedastic) $\Sigma^{(\mathrm{heterosc})} \in \mathbb{R}^{d \times d}$ is a diagonal matrix such that $\Sigma^{(\mathrm{heterosc})}_{jj} = 10$ for $1 \leq j \leq 5$, and $\Sigma^{(\mathrm{heterosc})}_{jj} = 0.1$ otherwise.
3. (Correlated) $\Sigma^{(\mathrm{corr})} \in \mathbb{R}^{d \times d}$ is a full matrix whose diagonal entries are equal to one and the off-diagonal terms are $\Sigma^{(\mathrm{corr})}_{jj'} = 1/\sqrt{|j - j'|}$ for $1 \leq j \neq j' \leq d$.

most times aRe used multiple tImes (RIver otter)

Comparison with existing literature



(a) Isotropic setting (b) Heteroscedastic setting (c) Correlated setting

Figure: Comparison of the empirical KL divergence between π_{data} and the generative distribution $\widehat{\pi}_N^{(\beta,\theta)}$ w.r.t. SGM for $\beta_{a^\star}$, the VPSDE model and the one with a cosine schedule, presented in Chen et al. (2023).

- Optimizing the noise schedule has an impact even with simple parametrization of the β scheduling.

most times aRe used multiple tImes (RIver otter)

Conclusion and perspectives

- $\rightsquigarrow$ Optimizing the noise schedule allows to optimize virtually all score-based samplers.
- $\rightsquigarrow$ Constants in the upper bound are crucial and should be taken care of.
- $\rightsquigarrow$ Important open problem: design π_{data} tailored (**Swift**) to π_∞ (then maybe use Langevin for the forward).

most times aRe used multiple tImes (RIver otter)