On the rate of convergence for score based diffusion models

Marta Gentiloni Silveri

CMAP, École Polytechnique LPSM, Sorbonne Université

Abstract

Score-based diffusion models (SGMs) are a new class of generative models that revolve around the estimation of the score function associated with a stochastic differential equation. Subsequent to its acquisition, the approximated score function is then harnessed to simulate the corresponding time-reversal process, ultimately enabling the generation of approximate data samples. The problem of establishing theoretical guarantees of convergence for diffusion models, that is to say the problem of estimating the distance between the output distribution and the sought data distribution, is still open. The main challenge is to quantify how the three sources of error entailed in each SGM - the time discretization error, the score approximation error and the initialization error - affect the quality of the returned samples. We present a novel method based on the mixture of ideas coming from stochastic control and functional inequalities that allows to derive simple, improved and sharp convergence bounds in KL applicable to any data distribution with finite Fisher information with respect to the standard Gaussian distribution. A joint work with Giovanni Conforti and Alain Durmus (Conforti et al. [2023]).

Score-Based Generative Models (SGMs)

Creating noise from data is easy, creating data from noise is generative modeling.
(Song et al. [2020])

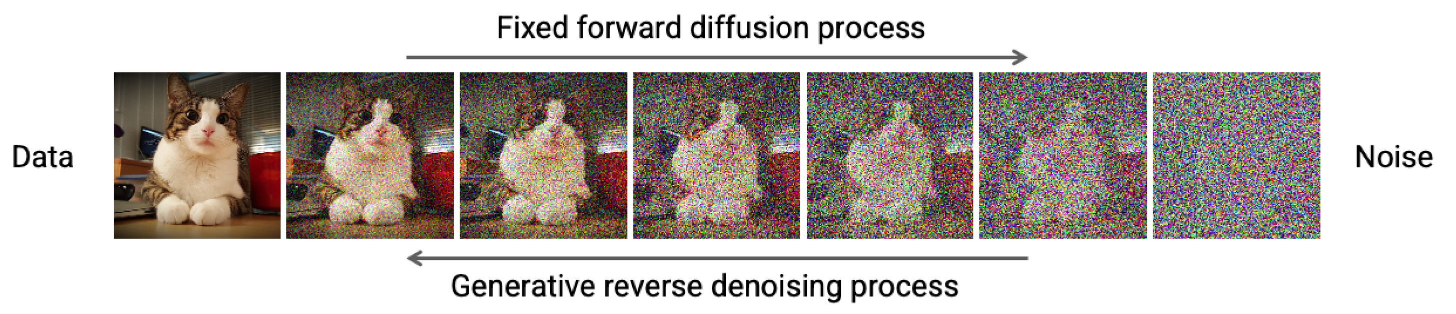
Goal of a SGM. Generate new samples similar to data ones $x \sim \mu_\star \in \mathcal{P}(\mathbb{R}^d)$.

Strategy. First, destruct progressively data by injecting noise:

$$\mathrm{d}\overrightarrow{X}_t = \mathbf{b}(\overrightarrow{X}_t)\mathrm{d}t + \boldsymbol{\Sigma}\mathrm{d}B_t\,, \quad t\in[0,T]\,, \quad \text{with} \quad \overrightarrow{X}_0 \sim \mu_\star\,, \qquad (1)$$

with $(\overrightarrow{X}_t)_{t\in[0,T]}$ d-dimensional ergodic diffusion associated to a Markov semi-group $(P_t)_{t\in[0,T]}$ with a unique stationary distribution μ_0. Second, reverse this process for sample generation, that is consider the solution to

$$\mathrm{d}\overleftarrow{X}_t = (-\mathbf{b}(\overleftarrow{X}_t) + \boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\nabla\log\overrightarrow{p}_{T-t}(\overleftarrow{X}_t))\mathrm{d}t + \boldsymbol{\Sigma}\mathrm{d}B_t\,, \quad t\in[0,T]\,, \quad \text{with} \quad \overleftarrow{X}_0 \sim \mu_\star P_T\,. \qquad (2)$$



Computational challenges to deal with

1. One cannot obtain i.i.d. samples from $\mu_\star P_T$;
2. The score of the forward process, $\nabla\log\overrightarrow{p}_{T-t}(x)$, which appears in (2), is intractable;
3. The continuous dynamics can not be simulated.

Solutions adopted.

1. Samples from the stationary distribution μ_0 of (1) are used instead;
2. An estimator $s_{\theta^\star}(t,x)$ is used instead. Among the neural networks $\{(t,x)\mapsto s_\theta(t,x)\}_{\theta\in\Theta}$, one picks the one that corresponds to the minimizer $\theta^\star$ of the score-matching objective
$$\theta \mapsto \int_0^T \mathbb{E}\left[\left\|s_\theta(t,\overrightarrow{X}_t) - \boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\nabla\log\overrightarrow{p}_t(\overrightarrow{X}_t)\right\|^2\right]\mathrm{d}t\,;$$
3. Discretizations schemes are used. Given a partition $\{0=t_0<t_1,\ldots<t_N=T\}$ of $[0,T]$ with meshes $\{h_k\}_k$ one considers the process $(X_t^{\mathrm{E}})_{t\in[0,T]}$ defined recursively on the intervals $[t_k,t_{k+1}]$ by
Euler-Maruyama (EM) discretization scheme:
$$\mathrm{d}X_t^{\mathrm{E}} = \{-\mathbf{b}(X_{t_k}^{\mathrm{E}}) + s_{\theta^\star}(T-t_k, X_{t_k}^{\mathrm{E}})\}\mathrm{d}t + \boldsymbol{\Sigma}\mathrm{d}B_t\,, \quad t\in[t_k,t_{k+1}]\,, \quad \text{with} \quad X_0^{\mathrm{E}}\sim\mu_0\,;$$
Euler Exponential Integrator (EI) scheme:
$$\mathrm{d}X_t^{\theta^\star} = \{-\mathbf{b}(X_t^{\theta^\star}) + s_{\theta^\star}(T-t_k, X_{t_k}^{\theta^\star})\}\mathrm{d}t + \boldsymbol{\Sigma}\mathrm{d}B_t, \quad t\in[t_k,t_{k+1}]\,, \quad \text{with} \quad X_0^{\theta^\star}\sim\mu_0\,. \qquad (3)$$

Resulting errors.

1. Initialization error;
2. Score approximation error;
3. Discretization error.

Main question

How do the various sources of error affect the quality of the returned samples?

Related literature:

Main strategies adopted up to now:

- assuming smoothness on the data distribution, compare $\mu_\star$ with the law at time T of the approximated backward process;
- introducing an early stopping rule, compare $\mu_\star P_\delta$ with the law at time $T-\delta$ of the approximated backward process.

Our setting

Consider the Ornstein–Uhlenbeck (OU) as forward process, so that (1) turns into

$$\mathrm{d}\overrightarrow{X}_t = -\overrightarrow{X}_t\mathrm{d}t + \sqrt{2}\mathrm{d}B_t\,, \quad t\in[0,T]\,, \quad \text{with} \quad \overrightarrow{X}_0\sim\mu_\star\,, \qquad (4)$$

$\mu_0 \equiv \gamma^d$ and (2) turns into

$$\mathrm{d}\overleftarrow{X}_t = (-\overleftarrow{X}_t + 2\nabla\log\tilde{p}_{T-t}(\overleftarrow{X}_t))\mathrm{d}t + \sqrt{2}\mathrm{d}B_t\,, \quad t\in[0,T]\,, \quad \text{with} \quad \overleftarrow{X}\sim\mu_\star P_T\,, \qquad (5)$$

where $\tilde{p}_t(x) := \overrightarrow{p}_t(x)/\gamma^d(x)$. Also, consider the EI as discretization scheme, so that (3) turns into

$$\mathrm{d}X_t^{\theta^\star} = (-X_t^{\theta^\star} + \tilde{s}_{\theta^\star}(T-t_k, X_{t_k}^{\theta^\star}))\mathrm{d}t + \sqrt{2}\mathrm{d}B_t\,, \quad t\in[t_k,t_{k+1}]\,, \quad \text{with} \quad X_0^{\theta^\star}\sim\gamma^d.$$

where $\tilde{s}_{\theta^\star}(t,x)$ is an estimator of $\tilde{p}_t(x)$ and $\{t_k\}_{k=1,\ldots,N}$ a partition of $[0,T]$ with meshes $h_k := t_k - t_{k-1}$.

Our contribution

Our assumption on the data distribution.

- **H1** $\mu_\star \ll \gamma^d$ and $\mu_\star$ has finite relative Fisher information against γ^d, *i.e.*
$$\mathscr{I}(\mu_\star|\gamma^d) = \int\left\|\nabla\log\left(\frac{\mathrm{d}\mu_\star}{\mathrm{d}\gamma^d}\right)\right\|^2\mathrm{d}\mu_\star < +\infty\,.$$

Our assumptions on the score approximation.

Either

- **H2** There exist $\varepsilon^2>0$ and $\theta^\star\in\mathbb{R}$ such that
$$\frac{1}{T}\sum_{k=0}^{N-1} h_{k+1}\mathbb{E}\left[\left\|\tilde{s}_{\theta^\star}(T-t_k,\overrightarrow{X}_{T-t_k}) - 2\nabla\log\tilde{p}_{T-t_k}(\overrightarrow{X}_{T-t_k})\right\|^2\right] \leq \varepsilon^2\,.$$

or

- **H3** There exist $\varepsilon^2>0$ and $\theta^\star\in\mathbb{R}$ such that, for any $k\in\{0,\ldots,N-1\}$,
$$\mathbb{E}\left[\left\|\tilde{s}_{\theta^\star}(T-t_k,\overrightarrow{X}_{T-t_k}) - 2\nabla\log\tilde{p}_{T-t_k}(\overrightarrow{X}_{T-t_k})\right\|^2\right] \leq \varepsilon^2\mathbb{E}\left[\left\|2\nabla\log\tilde{p}_{T-t_k}(\overrightarrow{X}_{T-t_k})\right\|^2\right]\,.$$

[Conforti et al., 2023, Theorem 2.1]

Let $T\geq 1, h\leq 1$ and assume **H1-H2**. Consider the EI scheme $(X_t^{\theta^\star})_{t\in[0,T]}$ with constant step size $h>0$. Denoting for any $t\in[0,T]$ by $p_t^{\theta^\star}$ the distribution of $X_t^{\theta^\star}$ we have that

$$\mathrm{KL}(\mu_\star|p_T^{\theta^\star}) \lesssim \mathrm{e}^{-2T}\mathrm{KL}(\mu_\star|\gamma^d) + \mathbf{C}(T,\varepsilon) + h\mathscr{I}(\mu_\star|\gamma^d)\,, \qquad (6)$$

where $\mathbf{C}(T,\varepsilon) = T\varepsilon^2$. Moreover, the above bound also holds if we replace the term $\mathrm{KL}(\mu_\star|\gamma^d)\mathrm{e}^{-2T}$ with $(\mathtt{M}_2^2+d)\mathrm{e}^{-T}$, where $\mathtt{M}_2^2$ is the second-order moment of $\mu_\star$.

Also, if instead of **H2**, **H3** holds, then (6) holds with $\mathbf{C}(T,\varepsilon) = \varepsilon^2\mathscr{I}(\mu_\star|\gamma^d)$.

Our method

Sketch of the proof.

- We interpreted the process $Y_t := 2\nabla\log\tilde{p}_{T-t}(\overleftarrow{X}_t)$ as the optimal drift in a stochastic control problem associated to (4);
- We studied its dynamics and derived the adjoint equation, that is
$$\mathrm{d}Y_t = Y_t\mathrm{d}t + \sqrt{2}\nabla Y_t\mathrm{d}B_t\,, \quad t\in[0,T]\,,$$
plus the exponential growth of $g(t) := \mathbb{E}[\|Y_t\|^2]$, *i.e.*the exponential decay of the Fisher along the semi-group associated to (4);
- We decomposed the KL as
$$\mathrm{KL}(\overleftarrow{P}|P^{\theta^\star}) = \mathrm{KL}(\overrightarrow{p}_T|\gamma^d) + \sum_{k=0}^{N-1}\frac{1}{4}\int_{kh}^{(k+1)h}\mathbb{E}\left[\left\|\tilde{s}_{\theta^\star}(T-kh,\overleftarrow{X}_{kh}) - Y_t\right\|^2\right]\mathrm{d}t\,,$$
and used all the information available to bound the (RHS).

Literature comparison

Table: Bounds on $\mathrm{KL}(\mu^\star|p_T^{\theta^\star})$ for the OU-based SGM stemming from EI with constant step-size.

Assumptions on the data	Related References	Error bound
H1 $\mathtt{M}_2^2<+\infty$ $\nabla\log\overrightarrow{p}_t$ $L-$ Lipschitz	[Chen et al., 2023, Theorem 2.1]	$(\mathtt{M}_2^2+d)\mathrm{e}^{-T}+T\varepsilon^2+dhL^2T$
H1 $\mathscr{I}(\mu_\star\|\gamma^d)\leq dL+\mathtt{M}_2^2$	[Conforti et al., 2023, Theorem 2.1]	$(\mathtt{M}_2^2+d)\mathrm{e}^{-T}+T\varepsilon^2+h(dL+\mathtt{M}_2^2)$

References

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pages 4735–4763. PMLR, 2023.

Giovanni Conforti, Alain Durmus, and Marta Gentiloni Silveri. Score diffusion models without early stopping: finite fisher information is all you need. *arXiv preprint arXiv:2308.12240*, 2023.

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.

Supervisors: Giovanni Conforti (Università degli studi di Padova, Alain Durmus (École Polytechnique), Maxime Sangnier (Sorbonne Université)

marta.gentiloni-silveri@polytechnique.edu