How Much Diffusion Time is Enough?

Giulio Franzese [1] **Simone Rossi** [1] **Lixuan Yang** [2] **Alessandro Finamore** [2] **Dario Rossi** [2] **Maurizio Filippone** [1]
Pietro Michiardi [1]

Abstract

Score-based diffusion models map noise into data using stochastic differential equations. While current practice advocates for a large T to ensure closeness to steady state, a smaller value of T should be preferred for a better approximation of the score-matching objective and computational efficiency. We conjecture, contrary to current belief and corroborated by numerical evidence, that the optimal diffusion times are smaller than current adoptions.

1. Introduction

Diffusion-based models (7; 10; 11; 14; 4; 2; 8) generate samples from an unknown density p_{data} by reversing a *diffusion process* which injects noise into the data. This diffusion process is a forward Stochastic Differential Equation (SDE)

$$\mathrm{d}\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, t)\mathrm{d}t + g(t)\mathrm{d}\boldsymbol{w}_t \quad \text{with} \quad \boldsymbol{x}_0 \sim p_{data}\,, \tag{1}$$

where $\boldsymbol{x}_t$ is a random variable at time t, $\boldsymbol{f}(\cdot, t)$ is the *drift term*, $g(\cdot)$ is the *diffusion term* and $\boldsymbol{w}_t$ is a *Wiener process*. We denote the time-varying probability density by $p(\boldsymbol{x}, t)$, by definition $p(\boldsymbol{x}, 0) = p_{data}(\boldsymbol{x})$, and the conditional on the initial condition $\boldsymbol{x}_0$ by $p(\boldsymbol{x}, t \,|\, \boldsymbol{x}_0)$. The forward SDE is usually considered for a sufficiently long *diffusion time* T as in principle, when $T \to \infty$, $p(\boldsymbol{x}, T)$ converges to Gaussian noise. Given initial condition $p(\boldsymbol{x}, T)$, the backward SDE (1)

$$\mathrm{d}\boldsymbol{x}_t = \left[-\boldsymbol{f}(\boldsymbol{x}_t, t') + g^2(t')\boldsymbol{\nabla}\log p(\boldsymbol{x}_t, t')\right]\mathrm{d}t + g(t')\mathrm{d}\boldsymbol{w}_t \quad t' \stackrel{\text{def}}{=} T - t, \tag{2}$$

after a *reverse diffusion time* T will be distributed as $p_{data}(\boldsymbol{x})$. Time varying density of Eq. (2) is denoted with $q(\boldsymbol{x}, t)$.

Practical considerations on diffusion time. In practice, diffusion models are challenging to work with (11). First, a direct access to the true *score* function $\boldsymbol{\nabla}\log p(\boldsymbol{x}_t, t)$ is required in the reverse diffusion is unavailable. This can be solved by approximating it with a parametric function $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$, which is trained using the following loss function,

$$\mathcal{L}(\boldsymbol{\theta}) = T\,\mathbb{E}_{t\sim\mathcal{U}(0,T)}\mathbb{E}_{\sim(1)}g^2(t)\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \boldsymbol{\nabla}\log p(\boldsymbol{x}_t, t \,|\, \boldsymbol{x}_0)\|^2\,, \tag{3}$$

where the notation $\mathbb{E}_{\sim(1)}$ means that the expectation is taken with respect to the random process $\boldsymbol{x}_t$ in Eq. (1). Considering affine drift, the term $p(\boldsymbol{x}_t, t \,|\, \boldsymbol{x}_0)$ is analytically known and normally distributed for all t (expression in Table 1, and in (12)). Intuitively, the estimation of the *score* is akin to a denoising objective, which operates in a challenging regime. Later we will quantify precisely the difficulty of learning the *score*, as a function of increasing diffusion times. Second, the noise distribution $p(\boldsymbol{x}, T)$ is analytically known only when the diffusion time is $T \to \infty$. The common solution is to replace $p(\boldsymbol{x}, T)$ with a simple distribution $p_{noise}(\boldsymbol{x})$ which, for the classes of SDEs we consider in this work, is a Gaussian distribution. Indeed, in the infinite diffusion time regime, it is possible to derive $p(\boldsymbol{x}, T \to \infty) = p_{noise}(\boldsymbol{x})$ analytically. We report in Table 1 the two main families of forward SDEs used in the literature, with the corresponding $p_{noise}(\boldsymbol{x})$.

[1]Department of Data Science, Eurecom, Biot, France [2]Huawei Technologies, Paris, France. Correspondence to: Giulio Franzese <giulio.franzese@eurecom.fr>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

Table 1: Two main families of diffusion processes

	Diffusion process	Marginal $p(\boldsymbol{x}_t, t \mid \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{m}, s\boldsymbol{I})$	$p_{noise}(\boldsymbol{x})$
Variance Exploding	$\alpha(t) = 0, g(t) = \sqrt{\frac{\mathrm{d}\sigma^2(t)}{\mathrm{d}t}}$	$\boldsymbol{m} = \boldsymbol{x}_0, s = \sigma^2(t) - \sigma^2(0)$	$\mathcal{N}(\mathbf{0}, \sigma^2(T) - \sigma^2(0)\boldsymbol{I})$
Variance Preserving	$\alpha(t) = \beta(t), g(t) = \sqrt{\beta(t)}$	$\boldsymbol{m} = c\boldsymbol{x}_0, s = 1 - c, c = e^{-\frac{1}{2}\int_0^t \beta(\mathrm{d}\tau)}$	$\mathcal{N}(\mathbf{0}, \boldsymbol{I})$

In the literature, the discrepancy between $p(\boldsymbol{x}, T)$ and $p_{noise}(\boldsymbol{x})$ has been neglected, under the informal assumption of a sufficiently large diffusion time. While this is a valid approach to simulate and generate samples, the reverse diffusion process starts from a different initial condition $q(\boldsymbol{x}, 0)$ and, as a consequence, it will converge to a solution $q(\boldsymbol{x}, T)$ that is different from the true $p_{data}(\boldsymbol{x})$. Later, we will expand on this error, but for illustration purposes Fig. 1 shows quantitatively this behavior for a simple 1D toy example $p_{data}(\boldsymbol{x}) = \pi\mathcal{N}(1, 0.1^2) + (1-\pi)\mathcal{N}(3, 0.5^2)$, with $\pi = 0.3$: when T is small, the distribution $p_{noise}(\boldsymbol{x})$ is very different from $p(\boldsymbol{x}, T)$ and samples from $q(\boldsymbol{x}, T)$ exhibit very low likelihood of being generated from $p_{data}(\boldsymbol{x})$.

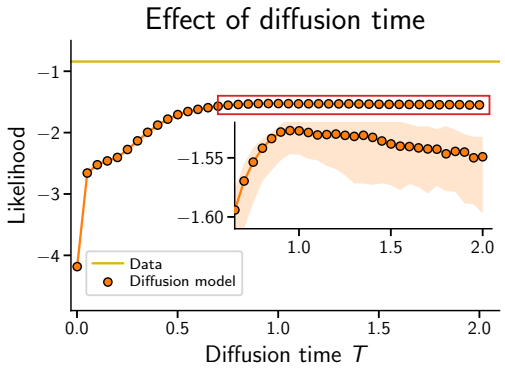

Figure 1: Effect of T on a toy model.

Crucially, Fig. 1 (zoomed region) illustrates an unknown behavior of diffusion models, unveiled in our analysis. In practice, there exists an optimal diffusion time that strikes right the balance between efficient *score* estimation, and sampling quality.

Contribution In § 2 we provide a new characterization of score-based diffusion models to obtain a formal understanding of the impact of the diffusion time T. We consider a decomposition of the evidence lower bound (ELBO), which emphasizes the roles of (i) the discrepancy between the "ending" distribution of the diffusion and the "starting" distribution of the reverse diffusion, and (ii) of the *score* matching objective. This allows us to investigate the existence of an optimal diffusion time $< \infty$, differently from current best practice for selecting T. In § 3 we provide experimental evidence of the described phenomenon.

2. A new ELBO decomposition and a tradeoff on diffusion time

The dynamics of a diffusion model can be studied through the lens of variational inference, bounding the (log-)likelihood using an evidence lower bound (ELBO) (3). Our interpretation emphasizes the two main factors affecting the quality of sample generation: an imperfect *score*, and a mismatch, measured in terms of the Kullback-Leibler (KL) divergence, between the noise distribution $p(\boldsymbol{x}, T)$ of the forward process and the distribution p_{noise} used to initialize the backward process.

2.1. The ELBO decomposition

By manipulating the $\mathcal{L}_{\text{ELBO}}$ derived in (3, Eq. (25)), we can write

$$\mathbb{E}_{p_{data}(\boldsymbol{x})} \log q(\boldsymbol{x}, T) \geq \mathcal{L}_{\text{ELBO}}(\boldsymbol{s}_{\boldsymbol{\theta}}, T) = \mathbb{E}_{\sim(1)} \log p_{noise}(\boldsymbol{x}_T) - I(\boldsymbol{s}_{\boldsymbol{\theta}}, T) + R(T), \tag{4}$$

where $R(T) = \frac{1}{2}\int_{t=0}^{T} \mathbb{E}_{\sim(1)} \left[g^2(t) \|\boldsymbol{\nabla} \log p(\boldsymbol{x}_t, t \mid \boldsymbol{x}_0)\|^2 - 2\boldsymbol{f}^\top(\boldsymbol{x}_t, t) \boldsymbol{\nabla} \log p(\boldsymbol{x}_t, t \mid \boldsymbol{x}_0) \right] \mathrm{d}t$, and $I(\boldsymbol{s}_{\boldsymbol{\theta}}, T) = \frac{1}{2}\int_{t=0}^{T} g^2(t) \mathbb{E}_{\sim(1)} \left[\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \boldsymbol{\nabla} \log p(\boldsymbol{x}_t, t \mid \boldsymbol{x}_0)\|^2 \right] \mathrm{d}t$. Note that $R(T)$ depends neither on $\boldsymbol{s}_{\boldsymbol{\theta}}$ nor on p_{noise}, while $I(\boldsymbol{s}_{\boldsymbol{\theta}}, T)$, or an equivalent reparameterization (3; 9, Eq. (1)), is used to learn the approximated *score*, by optimization of the parameters $\boldsymbol{\theta}$. It is then possible to show that

$$I(\boldsymbol{s}_{\boldsymbol{\theta}}, T) \geq \underbrace{I(\boldsymbol{\nabla} \log p, T)}_{\stackrel{\text{def}}{=} K(T)} = \frac{1}{2}\int_{t=0}^{T} g^2(t) \mathbb{E}_{\sim(1)} \left[\|\boldsymbol{\nabla} \log p(\boldsymbol{x}_t, t) - \boldsymbol{\nabla} \log p(\boldsymbol{x}_t, t \mid \boldsymbol{x}_0)\| \right]^2 \mathrm{d}t. \tag{5}$$

Consequently, we can rewrite $I(\boldsymbol{s}_{\boldsymbol{\theta}}, T)) = K(T) + \mathcal{G}(\boldsymbol{s}_{\boldsymbol{\theta}}, T)$, where $\mathcal{G}(\boldsymbol{s}_{\boldsymbol{\theta}}, T)$ is a positive term that we call the *gap* term, accounting for the practical case of an imperfect *score*, i.e. $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) \neq \boldsymbol{\nabla} \log p(\boldsymbol{x}_t, t)$. It also holds that

$$\mathbb{E}_{\sim(1)} \log p_{noise}(\boldsymbol{x}_T) = \int \left[\frac{\log p_{noise}(\boldsymbol{x}) p(\boldsymbol{x}, T)}{p(\boldsymbol{x}, T)} \right] p(\boldsymbol{x}, T) \mathrm{d}\boldsymbol{x} = \mathbb{E}_{\sim(1)} \log p(\boldsymbol{x}_T, T) - \mathrm{KL}\left[\log p(\boldsymbol{x}, T) \parallel p_{noise}(\boldsymbol{x}) \right]. \quad (6)$$

Therefore, we can rewrite the ELBO in Eq. (4) as

$$\mathbb{E}_{p_{data}(\boldsymbol{x})} \log q(\boldsymbol{x}, T) \geq -\mathrm{KL}\left[p(\boldsymbol{x}, T) \parallel p_{noise}(\boldsymbol{x}) \right] + \mathbb{E}_{\sim(1)} \log p(\boldsymbol{x}_T, T) - K(T) + R(T) - \mathcal{G}(\boldsymbol{s}_{\boldsymbol{\theta}}, T). \quad (7)$$

Before concluding our derivation it is necessary to introduce an important observation.

Proposition 1. *Given the stochastic dynamics defined in Eq.* (1)*, it holds that*

$$\mathbb{E}_{\sim(1)} \log p(\boldsymbol{x}_T, T) - K(T) + R(T) = \mathbb{E}_{p_{data}(\boldsymbol{x})} \log p_{data}(\boldsymbol{x}). \quad (8)$$

Intuitively, when $\boldsymbol{s}_{\boldsymbol{\theta}} = \nabla \log p$ (and consequently $I(\boldsymbol{s}_{\boldsymbol{\theta}}, T) = K(T)$), Eq. (4) is attained with equality. Moreover when $p_{noise}(\boldsymbol{x}) = p(\boldsymbol{x}, T)$, then $q(\boldsymbol{x}, T) = p_{data}(\boldsymbol{x})$. The formal justification of Proposition 1 is obtained by manipulating the results in (3) and the equality between $q(\boldsymbol{x}, t')$ and $p(\boldsymbol{x}, t)$ when the score estimation is exact and $q(\boldsymbol{x}, 0) = p(\boldsymbol{x}, T)$. Finally, we can now bound the value of $\mathbb{E}_{p_{data}(\boldsymbol{x})} \log q(\boldsymbol{x}, T)$ as

$$\mathbb{E}_{p_{data}(\boldsymbol{x})} \log q(\boldsymbol{x}, T) \geq \underbrace{\mathbb{E}_{p_{data}(\boldsymbol{x})} \log p_{data}(\boldsymbol{x}) - \mathcal{G}(\boldsymbol{s}_{\boldsymbol{\theta}}, T) - \mathrm{KL}\left[p(\boldsymbol{x}, T) \parallel p_{noise}(\boldsymbol{x}) \right]}_{\mathcal{L}_{\mathrm{ELBO}}(\boldsymbol{s}_{\boldsymbol{\theta}}, T)}. \quad (9)$$

Eq. (9) clearly emphasizes the roles of an approximate score function, through the gap term $\mathcal{G}(\cdot)$, and the discrepancy between the noise distribution of the forward process, and the initial distribution of the reverse process, through the KL term. In the ideal case of perfect *score* matching, the ELBO in Eq. (9) is attained with equality. If, in addition, the initial conditions for the reverse process are ideal, i.e. $q(\boldsymbol{x}, 0) = p(\boldsymbol{x}, T)$, then the results in (1) allow us to claim that $q(\boldsymbol{x}, T) = p_{data}(\boldsymbol{x})$.

2.2. Is there an optimal diffusion time?

While diffusion processes are generally studied for $T \to \infty$, for practical reasons, diffusion times in score-based models have been arbitrarily set to be "sufficiently large" in the literature. Here we conjecture the existence of an optimal diffusion time, which strikes the right balance between the gap $\mathcal{G}(\cdot)$ and the KL terms of the ELBO in Eq. (9).

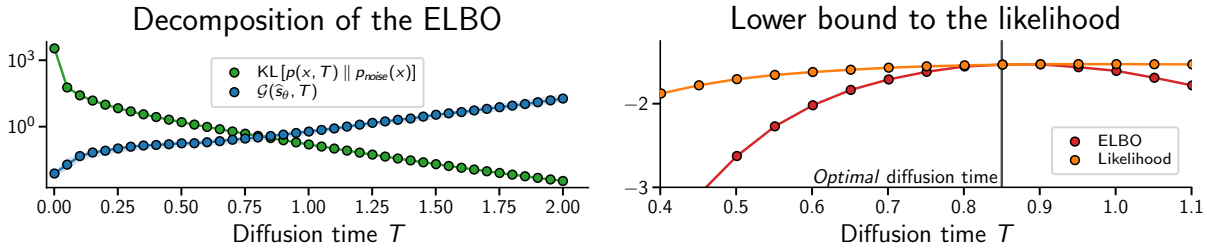

Figure 2: ELBO decomposition, ELBO and likelihood for a 1D toy model, as a function of diffusion time T. Tradeoff and optimality numerical results confirm our theory.

Empirically, we use Fig. 2 to illustrate this tradeoff through the lens of the same toy example we use in § 1. On the left, we show the ELBO decomposition. We can observe that $\mathcal{G}(\widehat{\boldsymbol{s}}_{\boldsymbol{\theta}}, T)$, the *gap* term obtained with the optimal set of parameters $\boldsymbol{\theta}$ for each T, is an increasing function of T, whereas the KL term is a decreasing function of T. Even in the simple case of a toy example, the tension between small and large values of T is clear. On the right, we show the values of the ELBO and of the likelihood as a function of T. We then verify the validity of our claims: the ELBO is neither maximized by an infinite diffusion time, nor by a "sufficiently large" value. Instead, there exists an optimal diffusion time $T^\star \approx 0.85$ which, for this example, is smaller than what is typically used in practical implementations, i.e. $T = 1.0$

3. Experiments

We now present numerical results on the MNIST and CIFAR10 datasets, to support our conjecture in § 2. We follow a standard experimental setup (8; 9; 3; 4): we use a standard U-Net architecture with time embeddings (2) and we report the log-likelihood in terms of bit per dimension (BPD) and the Fréchet Inception Distance (FID) scores (uniquely for CIFAR10). Although the FID score is a standard metric for ranking generative models, caution should be used against over-interpreting FID improvements (6). Similarly, while the theoretical properties of the models we consider are obtained through the lens of ELBO maximization, the log-likelihood measured in terms of BPD should be considered with care (13). Finally, we also report the number of neural function evaluations (NFE) for computing the relevant metrics. Training and evaluation is performed on a small cluster with 16 NVIDIA V100 GPUs. We considered Variance Preserving SDE with default β_0, β_1 parameter settings. When experimenting on CIFAR10 we considered the NCSN++ architecture as implemented in (11). Training of the score matching network has been carried out with the default set of optimizers and schedulers of (11), independently of the selected T. For the MNIST dataset we reduced the architecture by considering 64 features, ch_mult $= (1, 2)$ and attention resolutions equal to 8. The optimizer has been selected as the one in the CIFAR10 experiment but the warmup has been reduced to 1000 and the total number of iterations to 65000.

Exploring different diffusion times. We look for further empirical evidence of the existence of an optimal time. We shall focus on the baseline model (11), results are reported in Table 2. For MNIST, we observe how times $T = 0.6$ and $T = 1.0$ have comparable performance in terms of BPD, implying that any $T \geq 1.0$ is at best unnecessary and generally detrimental. Similarly, for CIFAR10, it is possible to notice that the best value of BPD is achieved for $T = 0.6$, outperforming all other values.

Table 2: Optimal T

Dataset	Time T	BPD (↓)	NFE (↓)
MNIST	1.0	1.16	300
	0.6	**1.16**	258
	0.4	1.25	235
	0.2	1.75	191
CIFAR10	1.0	3.09	221
	0.6	**3.07**	200
	0.4	3.09	187
	0.2	3.38	176

Dataset	$T = 1.0$	$T = 0.6$	$T = 0.4$	$T = 0.2$
FID	3.64	5.74	24.91	339.72
NFE	1000	600	400	200

Training and sampling efficiency Reducing T has the benefits of reducing training and sampling cost. For training, smaller times T allows to use simpler parametric score models. Sampling speed benefits are evident from Table 2. When considering the SDE version of the methods the number of sampling steps can decrease linearly with T, in accordance to theory (5), while retaining good BPD and FID scores. Similarly, although not in a linear fashion, the number of steps of the Ordinary Differential Equation (ODE) samplers can be reduced by using a smaller diffusion time T.

4. Conclusion

References

[1] B. D. Anderson. Reverse-Time Diffusion Equation Models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

[2] J. Ho, A. Jain, and P. Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[3] C.-W. Huang, J. H. Lim, and A. C. Courville. A Variational Perspective on Diffusion-Based Generative Models and Score Matching. In *Advances in Neural Information Processing Systems*, volume 34, pages 22863–22876. Curran Associates, Inc., 2021.

[4] D. Kingma, T. Salimans, B. Poole, and J. Ho. Variational Diffusion Models. In *Advances in Neural Information Processing Systems*, volume 34, pages 21696–21707. Curran Associates, Inc., 2021.

[5] P. E. Kloeden and E. Platen. A Survey of Numerical Methods for Stochastic Differential Equations. *Stochastic Hydrology and Hydraulics*, 3(3):155–178, 1989.

[6] T. Kynkäänniemi, T. Karras, M. Aittala, T. Aila, and J. Lehtinen. The Role of ImageNet Classes in Fréchet Inception Distance. *CoRR*, abs/2203.06026, 2022.

[7] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[8] J. Song, C. Meng, and S. Ermon. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*, 2021.

[9] Y. Song, C. Durkan, I. Murray, and S. Ermon. Maximum Likelihood Training of Score-Based Diffusion Models. In *Advances in Neural Information Processing Systems*, volume 34, pages 1415–1428. Curran Associates, Inc., 2021.

[10] Y. Song and S. Ermon. Generative Modeling by Estimating Gradients of the Data Distribution. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[11] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*, 2021.

[12] S. Särkkä and A. Solin. *Applied Stochastic Differential Equations*. Institute of Mathematical Statistics Textbooks. Cambridge University Press, 2019.

[13] L. Theis, A. van den Oord, and M. Bethge. A Note on the Evaluation of Generative Models. In Y. Bengio and Y. LeCun, editors, *International Conference on Learning Representations*, 2016.

[14] A. Vahdat, K. Kreis, and J. Kautz. Score-based Generative Modeling in Latent Space. In *Advances in Neural Information Processing Systems*, volume 34, pages 11287–11302. Curran Associates, Inc., 2021.