

Self-Regulating Annealing in Heavy-Tailed Diffusion Models

Keito Wakatsuki

Graduate School of Informatics
Kyoto University, Kyoto, Japan
wakatsuki.keito.63c@st.kyoto-u.ac.jp

Hideaki Shimazaki

Graduate School of Informatics
Kyoto University, Kyoto, Japan
h.shimazaki@i.kyoto-u.ac.jp

Abstract—Diffusion models have emerged as a leading framework for deep generative modeling. While the standard Gaussian formulation is theoretically convenient, its suitability for heavy-tailed datasets remains unclear. To address this, heavy-tailed diffusion models (HTDMs) [1] extend the standard formulation by replacing the Gaussian distribution with a Student’s t -distribution, thereby improving tail fidelity on heavy-tailed datasets. Although stochastic differential equation (SDE)-based sampling is possible in HTDMs, it has not been fully explored. In this paper, we propose an SDE-based sampler for HTDMs that explicitly incorporates a state-dependent diffusion coefficient. This state dependence naturally induces a self-regulating annealing mechanism by adaptively modulating the effective noise scale. We theoretically explore this mechanism and experimentally verify its necessity for reproducing samples from a heavy-tailed distribution.

Index Terms—Diffusion Model, Heavy-Tailed Diffusion Model, Student’s t -Distribution, Stochastic Differential Equation

I. INTRODUCTION

Diffusion models have emerged as a leading framework for deep generative modeling in recent years, particularly in image generation and other computer vision tasks [2], [3]. Denoising diffusion probabilistic models (DDPMs) [4] and their continuous-time formulations via stochastic differential equations (SDEs) and ordinary differential equations (ODEs) [5] provide a foundational framework and continue to attract substantial research interest.

In standard diffusion models, both the forward diffusion process and the reverse generative process are formulated using Gaussian transitions. While this Gaussian formulation yields analytically tractable marginals and posteriors, as well as simple training objectives, its suitability for heavy-tailed or outlier-prone datasets remains unclear. Motivated by this concern, recent studies have explored *non-Gaussian* diffusion models that replace the Gaussian formulation with heavy-tailed alternatives, such as approaches based on α -stable Lévy processes [6] and the Student’s t -distribution [1]. In particular, Pandey et al. [1] propose heavy-tailed diffusion models (HTDMs), a Student’s t -based extension that is readily compatible with standard diffusion models and requires only minimal modifications. Compared with standard diffusion models, HTDMs offer controllable tail behavior and better coverage of outlier regions.

In parallel, there has been growing interest in theoretically analyzing the generative dynamics of diffusion models from

a statistical-physics perspective [7], [8]. In particular, these dynamics have been shown to be theoretically connected to the memory retrieval dynamics of associative memory models [9], [10]. Associative memory models have also been extended using deformed exponential families, including the Student’s t -distribution [11]. This extension yields curved neural networks that exhibit rich and complex phenomena, such as explosive phase transitions with hysteresis and increased memory capacity. These phenomena arise from the stochastic dynamics of memory retrieval driven by a *self-regulating annealing mechanism*, which modulates memory retrieval via state-dependent changes in the effective temperature induced by higher-order statistical structure inherent to the model.

While similar dynamical properties may also arise in diffusion models extended with deformed exponential families, the effects of such deformations have not been fully explored. For HTDMs, Pandey et al. [1] derive both an ODE and an SDE from the Student’s t -based reverse kernel. However, the empirical evaluation in [1] primarily focuses on the ODE, leaving the SDE largely unexplored. To bridge this gap, we conduct a detailed analysis of the SDE for HTDMs and propose a novel SDE-based sampler. Our proposed sampler incorporates a state-dependent diffusion coefficient, which we argue is necessary to reproduce samples from a heavy-tailed distribution. Furthermore, we theoretically demonstrate that this state dependence induces the self-regulating annealing mechanism. The effective annealing temperature is automatically modulated based on the distance between the current and target states at each time step. This mechanism is closely related to the dynamics of memory retrieval in deformed associative memory models.

Our main contributions are summarized as follows:

- We propose an SDE-based sampler for HTDMs that incorporates a state-dependent diffusion coefficient.
- We demonstrate that this state dependence induces a self-regulating annealing mechanism.
- We empirically show that our SDE-based sampler improves tail fidelity on a synthetic dataset relative to a Gaussian baseline and an ablated SDE variant without the state-dependent diffusion coefficient.

Code availability: <https://github.com/keito93/self-regulating-annealing-in-htdms.git>

II. BACKGROUND

A. Diffusion Models

We briefly review the standard formulation of diffusion models as latent-variable models [4], [12]. Let $\mathbf{x}_0 \in \mathbb{R}^d$ denote an observed variable drawn from the data distribution p_{data} . The latent variables $\mathbf{x}_{\Delta t}, \mathbf{x}_{2\Delta t}, \dots, \mathbf{x}_{N\Delta t} \in \mathbb{R}^d$ are generated by a fixed diffusion process q that gradually corrupts $\mathbf{x}_0$ into pure noise. The generative model p_θ , parameterized by a neural network, defines a generative process that aims to reverse the diffusion process. When both q and p_θ are modeled as Markov chains (as in standard diffusion models), the negative evidence lower bound (negative ELBO) is given by

$$\begin{aligned} \mathcal{L}(\theta) = & -\mathbb{E}_{q(\mathbf{x}_{\Delta t}|\mathbf{x}_0)}[\ln p_\theta(\mathbf{x}_0 | \mathbf{x}_{\Delta t})] + \mathcal{D}_0(q(\mathbf{x}_T | \mathbf{x}_0) \| p(\mathbf{x}_T)) \\ & + \sum_{t=2\Delta t}^T \mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)}[\mathcal{D}_0(q(\mathbf{x}_{t-\Delta t} | \mathbf{x}_t, \mathbf{x}_0) \| p_\theta(\mathbf{x}_{t-\Delta t} | \mathbf{x}_t))], \end{aligned} \quad (1)$$

where $T := N\Delta t$, $t \in \{\Delta t, 2\Delta t, \dots, T\}$ denotes the discrete time step, and $\mathcal{D}_0$ denotes the Kullback–Leibler (KL) divergence. The second term in (1) matches the prior $p(\mathbf{x}_T)$ of the generative process to the marginal $q(\mathbf{x}_T | \mathbf{x}_0)$ of the diffusion process. The third term, which we refer to as the *denoising matching term*, requires the trainable reverse kernel $p_\theta(\mathbf{x}_{t-\Delta t} | \mathbf{x}_t)$ to approximate the true posterior $q(\mathbf{x}_{t-\Delta t} | \mathbf{x}_t, \mathbf{x}_0)$. In practice, this term serves as the training objective. To make the denoising matching term tractable, it is common to choose the forward process $q(\mathbf{x}_t | \mathbf{x}_{t-\Delta t})$ so that both the marginal $q(\mathbf{x}_t | \mathbf{x}_0)$ and the posterior $q(\mathbf{x}_{t-\Delta t} | \mathbf{x}_t, \mathbf{x}_0)$ can be obtained in closed form. If $q(\mathbf{x}_t | \mathbf{x}_{t-\Delta t})$ is a linear Gaussian transition, then the corresponding marginal and posterior are also Gaussian and analytically tractable. In this case, with an appropriate choice of the reverse kernel $p_\theta(\mathbf{x}_{t-\Delta t} | \mathbf{x}_t)$, the denoising matching term reduces to a mean-squared-error (MSE) loss.

After training, sampling starts from $\mathbf{x}_T \sim p(\mathbf{x}_T)$, and a synthetic sample $\mathbf{x}_0$ is obtained by iteratively applying the reverse kernel $p_\theta(\mathbf{x}_{t-\Delta t} | \mathbf{x}_t)$. In the continuous-time limit $\Delta t \rightarrow 0$, this discrete-time generative process converges to an SDE, which can be solved numerically to generate samples. While Gaussian diffusion models are highly successful in practice, it remains unclear whether the Gaussian formulation is well suited to heavy-tailed or outlier-prone datasets, which motivates the heavy-tailed extension discussed next.

B. Heavy-Tailed Diffusion Models

Pandey et al. [1] propose HTDMs and instantiate them as t -EDM, a Student’s t -based extension of the EDM design of Karras et al. [13]. We follow their construction and slightly reformulate it to align with our subsequent discussion.

They start from the same negative ELBO as in (1). The forward diffusion process is defined as an Nd -dimensional multivariate Student’s t -distribution over all time steps:

$$q(\mathbf{x}_{\Delta t}, \mathbf{x}_{2\Delta t}, \dots, \mathbf{x}_{N\Delta t} | \mathbf{x}_0) := \text{St}_{Nd}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu), \quad (2)$$

$$\boldsymbol{\mu} := [\mu_{\Delta t}, \mu_{2\Delta t}, \dots, \mu_{N\Delta t}]^\top \otimes \mathbf{x}_0, \quad (3)$$

$$\boldsymbol{\Sigma} := \tilde{\boldsymbol{\Sigma}} \otimes \mathbf{I}_d, \quad (4)$$

where $\nu > 2$ denotes the degrees of freedom and $\tilde{\boldsymbol{\Sigma}}$ is an $N \times N$ symmetric tridiagonal matrix whose diagonal entries are $(\tilde{\boldsymbol{\Sigma}})_{k,k} = \sigma_k^2 \Delta t$ for $k = 1, 2, \dots, N$, and whose first off-diagonal entries are $(\tilde{\boldsymbol{\Sigma}})_{k,k+1} = (\tilde{\boldsymbol{\Sigma}})_{k+1,k} = c_{(k+1)\Delta t}^2$ for $k = 1, 2, \dots, N-1$. $\mathbf{I}_d$ denotes the $d \times d$ identity matrix, and $\otimes$ denotes the Kronecker product. Note that under this definition, the diffusion process is non-Markovian.

A key property of this formulation is that the marginals and posteriors in the denoising matching term can be obtained in closed form by marginalization [14]:

$$q(\mathbf{x}_t | \mathbf{x}_0) = \text{St}_d(\mu_t \mathbf{x}_0, \sigma_t^2 \mathbf{I}_d, \nu), \quad (5)$$

$$q(\mathbf{x}_{t-\Delta t} | \mathbf{x}_t, \mathbf{x}_0) = \text{St}_d\left(\bar{\boldsymbol{\mu}}_t, \frac{\nu + \Delta t^2}{\nu + d} \bar{\sigma}_t^2 \mathbf{I}_d, \nu + d\right), \quad (6)$$

$$\bar{\boldsymbol{\mu}}_t := \frac{c_t^2}{\sigma_t^2} \mathbf{x}_t + \left(\mu_{t-\Delta t} - \frac{c_t^2}{\sigma_t^2} \mu_t\right) \mathbf{x}_0, \quad (7)$$

$$\bar{\sigma}_t^2 := \sigma_{t-\Delta t}^2 - \frac{c_t^4}{\sigma_t^2}, \quad (8)$$

$$\Delta_t^2 := \frac{1}{\sigma_t^2} \|\mathbf{x}_t - \mu_t \mathbf{x}_0\|^2, \quad (9)$$

where Δ_t^2 denotes the squared Mahalanobis distance. The schedule of μ_t is designed so that the marginal $q(\mathbf{x}_T | \mathbf{x}_0)$ approximately matches the prior $p(\mathbf{x}_T) = \text{St}_d(\mathbf{0}, \sigma_T^2 \mathbf{I}_d, \nu)$, and σ_t is chosen to increase monotonically with t .

Based on the posterior in (6), Pandey et al. [1] define the trainable reverse kernel $p_\theta(\mathbf{x}_{t-\Delta t} | \mathbf{x}_t)$ as a Student’s t -distribution. Since the KL divergence between Student’s t -distributions is intractable, they replace the KL divergence in the denoising matching term with a scaled Eguchi γ -power divergence [15], [16]. This scaled γ -power divergence is an extension of the KL divergence, to which it converges in the limit $\gamma \rightarrow 0$. Furthermore, to simplify the loss function, they omit the state-dependent factor $(\nu + \Delta_t^2)/(\nu + d)$ in the posterior covariance and approximate the posterior as

$$q(\mathbf{x}_{t-\Delta t} | \mathbf{x}_t, \mathbf{x}_0) \approx \text{St}_d(\bar{\boldsymbol{\mu}}_t, \bar{\sigma}_t^2 \mathbf{I}_d, \nu + d). \quad (10)$$

Using this simplified posterior, they parameterize the reverse kernel as

$$p_\theta(\mathbf{x}_{t-\Delta t} | \mathbf{x}_t) = \text{St}_d(\boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \bar{\sigma}_t^2 \mathbf{I}_d, \nu + d), \quad (11)$$

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) := \frac{c_t^2}{\sigma_t^2} \mathbf{x}_t + \left(\mu_{t-\Delta t} - \frac{c_t^2}{\sigma_t^2} \mu_t\right) \mathbf{D}_\theta(\mathbf{x}_t, \sigma_t). \quad (12)$$

Here, $\mathbf{D}_\theta(\mathbf{x}_t, \sigma_t) : \mathbb{R}^d \times \mathbb{R}_+ \rightarrow \mathbb{R}^d$ is a neural network that predicts $\mathbf{x}_0$ from $\mathbf{x}_t$, and aims to approximate $\mathbb{E}[\mathbf{x}_0 | \mathbf{x}_t]$. $\mathbf{D}_\theta(\mathbf{x}_t, \sigma_t)$ thus serves as a *denoiser*. With the scaled γ -power divergence and the above parameterization, the denoising matching term reduces to the familiar MSE loss:

$$\mathbb{E}_{t \sim \mathcal{U}(\{\Delta t, 2\Delta t, \dots, T\}), \mathbf{x}_0 \sim p_{\text{data}}, \mathbf{x}_t \sim q(\mathbf{x}_t | \mathbf{x}_0)} [\|\mathbf{x}_0 - \mathbf{D}_\theta(\mathbf{x}_t, \sigma_t)\|^2], \quad (13)$$

where $\mathcal{U}(\{\Delta t, 2\Delta t, \dots, T\})$ denotes the uniform distribution over discrete time steps.

C. Sampling of HTDMs

Given a trained denoiser, sampling starts from $\mathbf{x}_T \sim p(\mathbf{x}_T)$, and a synthetic sample $\mathbf{x}_0$ is obtained by iteratively applying the reverse kernel $p_\theta(\mathbf{x}_{t-\Delta t} | \mathbf{x}_t)$ defined in (11). From this discrete-time reverse kernel, Pandey et al. [1] derive both an ODE and an SDE for HTDMs in the continuous-time limit $\Delta t \rightarrow 0$. That is, for practical sampling, the generative process is defined and implemented as a Markov process, even though the diffusion process is non-Markovian. In our setting, their ODE is derived from a particular choice of c_t^2 in (8). With $c_t^2 = \sigma_t \sigma_{t-\Delta t}$, the variance of the reverse kernel, $\bar{\sigma}_t^2 = \sigma_{t-\Delta t}^2 - c_t^4/\sigma_t^2$, vanishes. Consequently, the reverse kernel becomes deterministic, and the continuous-time limit yields the following ODE:

$$\frac{d\mathbf{x}_t}{dt} = \frac{\dot{\mu}_t}{\mu_t} \mathbf{x}_t + \left[\frac{\dot{\sigma}_t}{\sigma_t} - \frac{\dot{\mu}_t}{\mu_t} \right] (\mathbf{x}_t - \mu_t \mathbf{D}_\theta(\mathbf{x}_t, \sigma_t)). \quad (14)$$

They also derive the SDE from the reverse kernel in (11), but do not further explore it as a practical sampler. In the next section, we focus on this SDE perspective and develop a practical SDE-based sampler that explicitly incorporates the state-dependent factor.

III. SELF-REGULATING ANNEALING IN HTDMs

A. SDE-Based Sampler for HTDMs

In this subsection, we derive an SDE-based sampler by taking the continuous-time limit $\Delta t \rightarrow 0$ of the reverse kernel in (11). In this limit, the generative process is described by a Wiener-driven SDE. We argue that reproducing heavy-tailed behavior in this SDE requires a state-dependent factor. To retain a non-vanishing noise term in the continuous-time limit, we choose c_t^2 in (8) as $c_t^2 = \sigma_{t-\Delta t}^2$. Under a smoothness assumption on σ_t , this choice yields

$$\bar{\sigma}_t^2 = \sigma_{t-\Delta t}^2 - \frac{c_t^4}{\sigma_t^2} = 2\sigma_t \dot{\sigma}_t \Delta t + \mathcal{O}(\Delta t^2). \quad (15)$$

Hence, the variance of the reverse kernel is of order Δt , so that the diffusion term remains in the continuous-time limit. Although the state-dependent factor $(\nu + \Delta t^2)/(\nu + d)$ appearing in the posterior covariance in (6) is omitted during training, we incorporate it during sampling in our SDE-based sampler. Its explicit incorporation distinguishes our approach from the previous study [1]. Using $c_t^2 = \sigma_{t-\Delta t}^2$ and a reparameterization of the Student's t -distribution, we can rewrite the reverse transitions as

$$\begin{aligned} \mathbf{x}_{t-\Delta t} = & \mathbf{x}_t + \left[\frac{\dot{\mu}_t}{\mu_t} \mathbf{x}_t - \left(\frac{\dot{\mu}_t}{\mu_t} - \frac{2\dot{\sigma}_t}{\sigma_t} \right) (\mathbf{x}_t - \mu_t \mathbf{D}_\theta(\mathbf{x}_t, \sigma_t)) \right] (-\Delta t) \\ & + \sqrt{\frac{\nu + \Delta t^2}{\nu + d}} \sqrt{2\sigma_t \dot{\sigma}_t} \sqrt{\Delta t} \frac{\mathbf{z}_t}{\sqrt{\kappa_t}}, \end{aligned} \quad (16)$$

where

$$\Delta_t^2 := \frac{1}{\sigma_t^2} \|\mathbf{x}_t - \mu_t \mathbb{E}[\mathbf{x}_0 | \mathbf{x}_t]\|^2 \approx \frac{1}{\sigma_t^2} \|\mathbf{x}_t - \mu_t \mathbf{D}_\theta(\mathbf{x}_t, \sigma_t)\|^2. \quad (17)$$

Here, $\mathbf{D}_\theta(\mathbf{x}_t, \sigma_t)$ is the denoiser trained to minimize the loss function in (13). Let $\mathbf{z}_t \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}_d)$ and $\kappa_t \sim \chi^2(\nu + d)/(\nu + d)$ be independent, where $\mathcal{N}_d(\mathbf{0}, \mathbf{I}_d)$ denotes the d -dimensional

standard Gaussian distribution and $\chi^2(\nu + d)$ denotes the chi-squared distribution with $\nu + d$ degrees of freedom. Then, $\mathbf{z}_t/\sqrt{\kappa_t} \sim \text{St}_d(\mathbf{0}, \mathbf{I}_d, \nu + d)$.

We next show that in the continuous-time limit $\Delta t \rightarrow 0$, the discrete-time update in (16) converges to an Itô SDE driven by a Wiener process. To this end, we consider a general forward-time update with Student's t noise:

$$\begin{aligned} \Delta \mathbf{x}_t &:= \mathbf{x}_{t+\Delta t} - \mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t) \Delta t + g(\mathbf{x}_t, t) \sqrt{\Delta t} \frac{\mathbf{z}_t}{\sqrt{\kappa_t}}, \quad (18) \\ \mathbf{z}_t &\sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}_d), \quad \kappa_t = \frac{\xi_t}{\nu + d}, \quad \xi_t \sim \chi^2(\nu + d), \quad (19) \end{aligned}$$

where $\mathbf{f}(\mathbf{x}_t, t) \in \mathbb{R}^d$ is a d -dimensional vector-valued function and $g(\mathbf{x}_t, t) \in \mathbb{R}$ is a scalar-valued function. The random variables $\mathbf{z}_t$ and κ_t are independent, and the pairs $(\mathbf{z}_t, \kappa_t)$ are i.i.d. over t .

Proposition 1 (Gaussian limit of Student's t increments). *Assume $\nu > 2$. For fixed $\mathbf{x}_t = \mathbf{x}$ and t , the conditional distribution of $\Delta \mathbf{x}_t$ is approximated by a Gaussian distribution to first order in Δt .*

$$p(\Delta \mathbf{x}_t | \mathbf{x}_t = \mathbf{x}) \approx \mathcal{N}_d \left(\mathbf{f}(\mathbf{x}, t) \Delta t, g^2(\mathbf{x}, t) \Delta t \frac{\nu + d}{\nu + d - 2} \mathbf{I}_d \right). \quad (20)$$

Consequently, in the continuous-time limit $\Delta t \rightarrow 0$, the resulting dynamics are described by an Itô SDE driven by a d -dimensional Wiener process:

$$d\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t) dt + g(\mathbf{x}_t, t) \sqrt{\frac{\nu + d}{\nu + d - 2}} d\mathbf{w}_t. \quad (21)$$

Proof. Fix $(\mathbf{x}, t)$ and set $\mathbf{f} := \mathbf{f}(\mathbf{x}, t)$, $g := g(\mathbf{x}, t)$. Define the characteristic function of $\Delta \mathbf{x}_t$ conditioned on $\mathbf{x}_t = \mathbf{x}$ as

$$\phi(\mathbf{u} | \mathbf{x}) := \mathbb{E}_{\mathbf{z}_t, \kappa_t} \left[e^{i\mathbf{u}^\top \Delta \mathbf{x}_t} | \mathbf{x}_t = \mathbf{x} \right]. \quad (22)$$

Using $\Delta \mathbf{x}_t = \mathbf{f} \Delta t + g \sqrt{\Delta t} \frac{\mathbf{z}_t}{\sqrt{\kappa_t}}$ and $\mathbf{z}_t \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}_d)$, we first take the conditional expectation over $\mathbf{z}_t$ given κ_t :

$$\begin{aligned} \phi(\mathbf{u} | \mathbf{x}) &= e^{i\mathbf{u}^\top \mathbf{f} \Delta t} \mathbb{E}_{\kappa_t} \left[\exp \left(-\frac{g^2 \Delta t}{2\kappa_t} \|\mathbf{u}\|^2 \right) \right] \\ &= 1 + i\mathbf{u}^\top \mathbf{f} \Delta t - \frac{1}{2} g^2 \Delta t \mathbb{E}[\kappa_t^{-1}] \|\mathbf{u}\|^2 + o(\Delta t). \end{aligned} \quad (23)$$

Since $\kappa_t = \xi_t/(\nu + d)$ and $\xi_t \sim \chi^2(\nu + d)$,

$$\mathbb{E}[\kappa_t^{-1}] = (\nu + d) \mathbb{E}[\xi_t^{-1}] = \frac{\nu + d}{\nu + d - 2}. \quad (24)$$

We therefore have

$$\begin{aligned} \phi(\mathbf{u} | \mathbf{x}) &= 1 + i\mathbf{u}^\top \mathbf{f} \Delta t - \frac{1}{2} \mathbf{u}^\top \left(g^2 \Delta t \frac{\nu + d}{\nu + d - 2} \mathbf{I}_d \right) \mathbf{u} + o(\Delta t), \end{aligned} \quad (25)$$

which is the characteristic function of the Gaussian distribution in (20) up to $o(\Delta t)$. $\square$

Applying Proposition 1 to the reverse-time update in (16), and noting that sampling proceeds backward in time from T to 0, we obtain the following reverse-time SDE:

$$d\mathbf{x}_t = \left[\frac{\dot{\mu}_t}{\mu_t} \mathbf{x}_t - \left(\frac{\dot{\mu}_t}{\mu_t} - \frac{2\dot{\sigma}_t}{\sigma_t} \right) (\mathbf{x}_t - \mu_t \mathbf{D}_\theta(\mathbf{x}_t, \sigma_t)) \right] dt + \alpha(\Delta_t^2(\mathbf{x}_t), t) \sqrt{2\sigma_t \dot{\sigma}_t} d\mathbf{w}_t, \quad (26)$$

$$\alpha(\Delta_t^2(\mathbf{x}_t), t) := \sqrt{\frac{\nu + \Delta_t^2}{\nu + d - 2}}. \quad (27)$$

We refer to $\alpha(\Delta_t^2(\mathbf{x}_t), t)$ as the state-dependent coefficient. We can generate samples by numerically solving this SDE from $t = T$ to $t = 0$, starting from $\mathbf{x}_T \sim p(\mathbf{x}_T) = \text{St}_d(\mathbf{0}, \sigma_T^2 \mathbf{I}_d, \nu)$.

For example, choosing $\mu_t = 1$ and $\sigma_t = \sigma\sqrt{t}$ yields

$$d\mathbf{x}_t = \frac{\mathbf{x}_t - \mathbf{D}_\theta(\mathbf{x}_t, \sigma_t)}{t} dt + \sigma \alpha(\Delta_t^2(\mathbf{x}_t), t) d\mathbf{w}_t, \quad (28)$$

$$\Delta_t^2 = \frac{1}{\sigma^2 t} \|\mathbf{x}_t - \mathbf{D}_\theta(\mathbf{x}_t, \sigma_t)\|^2. \quad (29)$$

This parameter choice corresponds to the variance-exploding (VE) setting in HTDMs. In the Gaussian limit $\nu \rightarrow \infty$ of the Student's t -distribution, the state-dependent coefficient converges to 1, and the SDE reduces to the reverse-time VE-SDE in [5] with constant diffusion coefficient σ . Algorithm 1 summarizes a discretized VE-SDE-based sampler for HTDMs. In implementation, we index the states by the noise level σ rather than time: we denote by $\mathbf{x}_i$ the state at noise level σ_i (with $\sigma_N = \sigma_{\max}$ and $\sigma_1 = \sigma_{\min}$), and set $\mathbf{x}_0$ to be the final denoised output.

Algorithm 1 SDE-based sampler ($\mu_t = 1, \sigma_t = \sigma\sqrt{t}$)

Require: Denoiser $\mathbf{D}_\theta(\mathbf{x}, \sigma)$, degrees of freedom ν , dimension d , number of steps N , noise-level sequence $\{\sigma_i\}_{i=1}^N$ ($\sigma_N = \sigma_{\max} > \dots > \sigma_1 = \sigma_{\min}$)

- 1: Sample $\mathbf{x}_N \sim \text{St}_d(\mathbf{0}, \sigma_N^2 \mathbf{I}_d, \nu)$ ▷ Initialization
 - 2: **for** $i = N$ **down to** 2 **do**
 - 3: $\mathbf{d}_i \leftarrow \frac{2(\mathbf{x}_i - \mathbf{D}_\theta(\mathbf{x}_i, \sigma_i))}{\sigma_i}$ ▷ Drift term
 - 4: $\Delta_i^2 \leftarrow \frac{\|\mathbf{x}_i - \mathbf{D}_\theta(\mathbf{x}_i, \sigma_i)\|^2}{\sigma_i^2}$
 - 5: $\alpha_i \leftarrow \sqrt{\frac{\nu + \Delta_i^2}{\nu + d - 2}}$ ▷ State-dependent coefficient
 - 6: Sample $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$
 - 7: $\mathbf{x}_{i-1} \leftarrow \mathbf{x}_i + \mathbf{d}_i (\sigma_{i-1} - \sigma_i) + \alpha_i \sqrt{\sigma_i^2 - \sigma_{i-1}^2} \mathbf{z}_i$
 - 8: **end for**
 - 9: $\mathbf{x}_0 \leftarrow \mathbf{D}_\theta(\mathbf{x}_1, \sigma_1)$ ▷ Final denoising
 - 10: **return** $\mathbf{x}_0$
-

B. Self-Regulating Annealing in the SDE-Based Sampler

In this subsection, we analyze our proposed VE-SDE-based sampler in (28) and elucidate its self-regulating annealing mechanism. We assume that the denoiser $\mathbf{D}_\theta(\mathbf{x}_t, \sigma_t)$ has been trained to accurately approximate $\mathbb{E}[\mathbf{x}_0 | \mathbf{x}_t]$. In other words, given a noisy state $\mathbf{x}_t$, the denoiser estimates a corresponding target (denoised) state. If we ignore the stochastic term in (28) and consider only the drift, the dynamics are driven toward this

target estimate. In particular, the current state is consistent with the target estimate when it satisfies the following fixed-point equation:

$$\mathbf{x}_t - \mathbf{D}_\theta(\mathbf{x}_t, \sigma_t) = \mathbf{0}. \quad (30)$$

On the other hand, the diffusion coefficient contains $\Delta_t^2 \propto \|\mathbf{x}_t - \mathbf{D}_\theta(\mathbf{x}_t, \sigma_t)\|^2$, which measures the distance between the current state and the target estimate at time t . Hence, when the current state is far from the target estimate, the diffusion coefficient increases, injecting stronger noise and promoting broader exploration; when it is close, the diffusion coefficient decreases, leading to milder noise and more gradual exploration. Equivalently, this mechanism can be interpreted as a state-dependent modulation of the effective annealing temperature, which we refer to as a *self-regulating annealing mechanism* [11].

We next demonstrate the self-regulating annealing mechanism in a setting where the SDE dynamics can be analyzed explicitly. Consider the symmetric two-point data distribution $p_{\text{data}}(\mathbf{x}_0) = \frac{1}{2}\delta(\mathbf{x}_0 - \mathbf{a}) + \frac{1}{2}\delta(\mathbf{x}_0 + \mathbf{a})$, for some fixed $\mathbf{a} \in \mathbb{R}^d$. In this case, $\mathbb{E}[\mathbf{x}_0 | \mathbf{x}_t]$ can be obtained in closed form as

$$\mathbb{E}[\mathbf{x}_0 | \mathbf{x}_t] = \mathbf{a} \tanh_\gamma(\beta_t(\mathbf{x}_t) \mathbf{a}^\top \mathbf{x}_t), \quad (31)$$

$$\beta_t(\mathbf{x}_t) := \frac{\nu + d}{\nu \sigma^2 t} \cdot \frac{1}{1 + \frac{1}{\nu \sigma^2 t} (\|\mathbf{x}_t\|^2 + \|\mathbf{a}\|^2)}, \quad (32)$$

where $\gamma := -\frac{2}{\nu+d}$. Here, $\tanh_\gamma(x)$ denotes the deformed hyperbolic tangent defined as

$$\tanh_\gamma(x) := \frac{(1 + \gamma x)^{\frac{1}{\gamma}} - (1 - \gamma x)^{\frac{1}{\gamma}}}{(1 + \gamma x)^{\frac{1}{\gamma}} + (1 - \gamma x)^{\frac{1}{\gamma}}} \quad \text{for } |\gamma x| < 1, \quad (33)$$

which converges to $\tanh(x)$ as $\gamma \rightarrow 0$ (equivalently, as $\nu \rightarrow \infty$). Assuming an ideal denoiser $\mathbf{D}_\theta(\mathbf{x}_t, \sigma_t) = \mathbb{E}[\mathbf{x}_0 | \mathbf{x}_t]$ under this data distribution, the SDE in (28) can be written as

$$d\mathbf{x}_t = \frac{\mathbf{x}_t - \mathbf{a} \tanh_\gamma(\beta_t(\mathbf{x}_t) \mathbf{a}^\top \mathbf{x}_t)}{t} dt + \sigma \alpha(\Delta_t^2(\mathbf{x}_t), t) d\mathbf{w}_t, \quad (34)$$

$$\Delta_t^2 = \frac{1}{\sigma^2 t} \|\mathbf{x}_t - \mathbf{a} \tanh_\gamma(\beta_t(\mathbf{x}_t) \mathbf{a}^\top \mathbf{x}_t)\|^2. \quad (35)$$

Figure 1 summarizes the key behavior of (34) in the one-dimensional setting with $a = 10$, $\nu = 3$, and $\sigma = 1$. Figure 1a plots $y = x$ and $y = a \tanh_\gamma(\beta_t(x) a x)$ for a representative value of $t \in [0, 1]$. Their intersections correspond to the fixed points (zeros of the drift), which satisfy the following fixed-point equation:

$$x = a \tanh_\gamma(\beta_t(x) a x). \quad (36)$$

This equation is analogous to the self-consistent equations in statistical physics, and can be interpreted as the HTDM counterpart of the result in [7], [8]. In the one-dimensional case, Δ_t^2 in (34) corresponds to the squared vertical gap between the two curves in Figure 1a, scaled by the noise level. Figure 1b shows sample trajectories of (34) over $t \in [0, 1]$ (solid) and those of an ablated baseline with $\alpha(\Delta_t^2(\mathbf{x}_t), t) \equiv 1$ (dashed), using the same Wiener increments for a fair comparison. Since $\alpha(\Delta_t^2(\mathbf{x}_t), t)$ increases with Δ_t^2 , the trajectories exhibit broader exploration when the state is far from the fixed points.

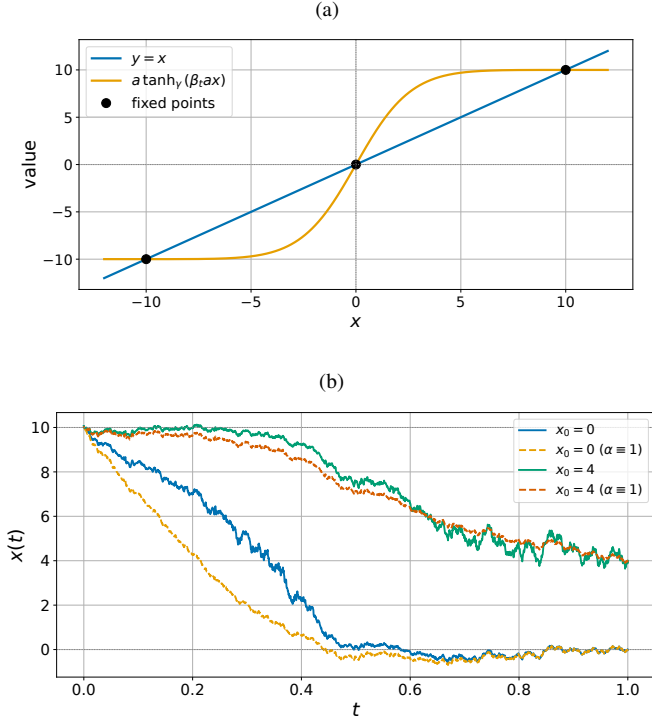


Fig. 1. Self-regulating annealing mechanism in (34) for the symmetric two-point data distribution in the one-dimensional setting ($a = 10, \nu = 3, \sigma = 1$). (a) Plots of $y = x$ and $y = a \tanh_{\gamma}(\beta_t(x) a x)$ for a representative value of $t \in [0, 1]$; their intersections correspond to the fixed points (zeros of the drift) satisfying (36). The curves change only slightly over $t \in [0, 1]$; hence, the curve at $t = 1$ is shown. (b) Sample trajectories over $t \in [0, 1]$: (34) (solid) versus an ablated baseline with $\alpha(\Delta_t^2(\mathbf{x}_t), t) \equiv 1$ (dashed), using the same Wiener increments for a fair comparison.

IV. EXPERIMENTS

We evaluate generative performance on a synthetic heavy-tailed dataset. To directly compare with the theoretical tail behavior, we use a one-dimensional Student’s t -distribution as the data distribution:

$$p_{\text{data}}(x) = \text{St}_1(x \mid 0, 1, 3). \quad (37)$$

We evaluate the performance of the following four samplers:

- 1) **VE-SDE (baseline)**: a Gaussian VE-SDE sampler for EDM [5], [13].
- 2) **t -ODE**: an ODE-based sampler for t -EDM [1].
- 3) **t -SDE**: our proposed SDE-based sampler for t -EDM that explicitly incorporates the state-dependent coefficient $\alpha(\Delta_t^2(\mathbf{x}_t), t)$.
- 4) **Ablated t -SDE**: an ablated variant of our SDE-based sampler for t -EDM in which $\alpha(\Delta_t^2(\mathbf{x}_t), t) \equiv 1$.

We focus on tail fidelity, using quantile–quantile (Q–Q) plots for qualitative evaluation and extreme-tail probabilities for quantitative comparison with the ground truth.

A. Experimental Setup

1) **Dataset**: We generate 10^6 i.i.d. samples from p_{data} for training and another 10^6 i.i.d. samples for testing, and apply z -score normalization.

2) **Denoiser Architecture**: We use the same denoiser architecture for all methods: a fully connected neural network with two hidden layers of width 128.

3) **Preconditioning and Training**: We follow the preconditioning and training settings in Table 7 of [1]: EDM settings for VE-SDE and t -EDM settings for t -ODE and both t -SDE variants. We train both EDM and t -EDM on a fixed budget of 10^7 samples and use the final checkpoint.

4) **Sampling**: VE-SDE and the two t -SDE variants are all based on the SDE in (28). For VE-SDE, we use the Gaussian limit ($\nu \rightarrow \infty$). The t -SDE retains the state-dependent coefficient $\alpha(\Delta_t^2(\mathbf{x}_t), t)$, whereas the ablated t -SDE sets it to 1. The t -ODE is based on the ODE in (14) under the VE parameterization, i.e., $\mu_t = 1$ and $\sigma_t = \sigma\sqrt{t}$. We choose the discrete noise levels $\{\sigma_i\}_{i=1}^N$ according to the EDM schedule of [13]. We solve the ODE using Heun’s method and the SDE using the Euler–Maruyama method.

5) **Hyperparameters**: Unless otherwise noted, we follow the hyperparameter settings in Table 7 of [1]. When training t -EDM, we set $\nu = 3$. For our SDE-based sampler, ν appears explicitly in both the prior and state-dependent coefficient. We therefore treat it as a sampling hyperparameter and sweep $\nu \in \{2.5, 3, 3.5, 4\}$, ultimately using $\nu = 2.5$ based on the evaluation criteria described below. Since the number of function evaluations (NFE) differs between Heun’s method and the Euler–Maruyama method, we choose the ODE and SDE step counts to match their NFE budgets, using 64 ODE steps and 128 SDE steps.

B. Evaluation

We first evaluate overall sample quality by computing the 1-Wasserstein distance W_1 from 10^6 generated and 10^6 test samples. We next assess tail fidelity using (i) Q–Q plots, which visualize the alignment of empirical quantiles between generated and test samples, and (ii) extreme-tail probabilities in the raw data space. For Q–Q plots, we use 10^6 generated samples and 10^6 test samples. We compute empirical quantiles on a uniform grid of 10^5 quantile levels in $[10^{-4}, 1 - 10^{-4}]$ to prevent extreme quantiles from distorting the axis range. This trimming is only for visualization; we still evaluate the extreme tails quantitatively using extreme-tail probabilities.

For extreme-tail probabilities, we measure the empirical rate of samples exceeding a fixed two-sided threshold u . This comparison is performed in the raw data space. We set u to the theoretical two-sided 0.999 quantile of the ground-truth distribution, i.e., $u := u_{0.999} = 12.924$, so that $\mathbb{P}_{x \sim p_{\text{data}}}(|x| > u) = 10^{-3}$. For each method, we generate 10^7 samples and split them into 10 disjoint batches of $M = 10^6$ samples. Within each batch, we estimate $\mathbb{P}(|x| > u)$ by the empirical exceedance rate $\hat{p} = \frac{1}{M} \sum_{i=1}^M \mathbf{1}[|x_i| > u]$, where $\mathbf{1}[\cdot]$ denotes the indicator function. We report the mean and standard deviation of $\hat{p}$ across batches, as well as the relative error with respect to the ground truth.

C. Results

Figure 2 shows the Q–Q plots of the four samplers. The VE-SDE and the ablated t -SDE deviate noticeably from the

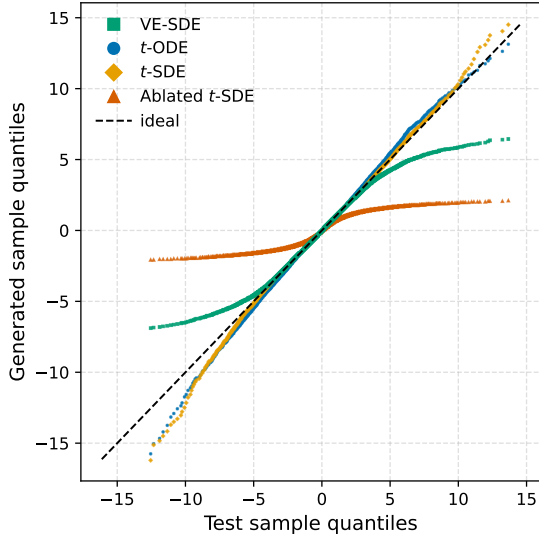


Fig. 2. Q-Q plots comparing generated and test samples for the four samplers.

TABLE I
OVERALL ACCURACY ($\mathcal{W}_1$) AND HEAVY-TAIL FIDELITY AT $u_{0.999}$.
EXTREME-TAIL PROBABILITIES ARE REPORTED AS MEAN $\pm$ STD ACROSS
BATCHES. WE DEFINE THE RELATIVE ERROR AS
 $\text{Rel. Err.} := (\hat{p} - 10^{-3})/10^{-3}$.

Method	$\mathcal{W}_1$	$\mathbb{P}(X > u_{0.999})$	Rel. Err.
VE-SDE	0.0287	$(5.630 \pm 0.689) \times 10^{-5}$	-0.9437
t -ODE	0.0202	$(1.307 \pm 0.029) \times 10^{-3}$	0.3073
t -SDE	0.0162	$(1.233 \pm 0.030) \times 10^{-3}$	0.2325
Ablated t -SDE	0.2233	0	-1

ideal $y = x$ line in the tail region, indicating poor tail fidelity. In contrast, t -ODE and t -SDE closely follow the ideal line, demonstrating superior tail fidelity.

Table I shows that overall sample quality is broadly comparable across methods, except for the ablated t -SDE, which exhibits a markedly larger $\mathcal{W}_1$. In contrast, tail fidelity differs substantially across methods: the t -ODE and t -SDE reproduce the correct order of the extreme-tail probability (10^{-3}), whereas the VE-SDE severely underestimates this probability and the ablated t -SDE reduces it to zero.

V. CONCLUSIONS

We revisit HTDMs from an SDE perspective and propose an SDE-based sampler. This sampler is driven by a Wiener process and explicitly incorporates the state-dependent diffusion coefficient induced by the Student’s t posterior. Owing to the state dependence of the diffusion coefficient, the generative dynamics exhibit a self-regulating annealing mechanism: the effective noise scale increases when the current state is far from the denoiser’s target estimate and decreases as the state approaches it. Experiments on a synthetic dataset confirm that this state dependence is necessary not only for faithfully reproducing the heavy tails but also for improving overall sample quality. As a result, our SDE-based sampler improves

tail fidelity relative to a Gaussian baseline and achieves performance comparable to the existing t -ODE-based sampler.

A. Limitations and Future Work

Our experimental evaluation is currently limited to a synthetic one-dimensional setting, and the degrees-of-freedom parameter ν used in sampling is selected through a small sweep. Extending both the analysis and the empirical validation to higher-dimensional benchmarks (e.g., images) and developing principled methods for selecting ν are promising directions for future work.

ACKNOWLEDGEMENTS

This work is supported by JST SPRING, Grant Number A94251400049 and JSPS KAKENHI, Grant Numbers JP 24K21518, 25K03085.

REFERENCES

- [1] K. Pandey, J. Pathak, Y. Xu, S. Mandt, M. Pritchard, A. Vahdat, and M. Mardani, “Heavy-tailed diffusion models,” arXiv:2410.14171, 2024, arXiv preprint.
- [2] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep unsupervised learning using nonequilibrium thermodynamics,” in *International conference on machine learning*. pmlr, 2015, pp. 2256–2265.
- [3] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, “Diffusion models: A comprehensive survey of methods and applications,” *ACM computing surveys*, vol. 56, no. 4, pp. 1–39, 2023.
- [4] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
- [5] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, “Score-based generative modeling through stochastic differential equations,” *arXiv preprint arXiv:2011.13456*, 2020.
- [6] E. B. Yoon, K. Park, S. Kim, and S. Lim, “Score-based generative models with lévy processes,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 40 694–40 707, 2023.
- [7] L. Ambrogioni, “The statistical thermodynamics of generative diffusion models: Phase transitions, symmetry breaking and critical instability,” *arXiv preprint arXiv:2310.17467*, 2023.
- [8] G. Raya and L. Ambrogioni, “Spontaneous symmetry breaking in generative diffusion models,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2024, no. 10, p. 104025, 2024.
- [9] L. Ambrogioni, “In search of dispersed memories: Generative diffusion models are associative memory networks,” *Entropy*, vol. 26, no. 5, p. 381, 2024.
- [10] B. Hoover, H. Strobelt, D. Krotov, J. Hoffman, Z. Kira, and D. H. Chau, “Memory in plain sight: Surveying the uncanny resemblances of associative memories and diffusion models,” *arXiv preprint arXiv:2309.16750*, 2023.
- [11] M. Aguilera, P. A. Morales, F. E. Rosas, and H. Shimazaki, “Explosive neural networks via higher-order interactions in curved statistical manifolds,” *Nature Communications*, vol. 16, no. 1, p. 6511, 2025.
- [12] C. Luo, “Understanding diffusion models: A unified perspective,” *arXiv preprint arXiv:2208.11970*, 2022.
- [13] T. Karras, M. Aittala, T. Aila, and S. Laine, “Elucidating the design space of diffusion-based generative models,” *Advances in neural information processing systems*, vol. 35, pp. 26 565–26 577, 2022.
- [14] P. Ding, “On the conditional distribution of the multivariate t distribution,” *The American Statistician*, vol. 70, no. 3, pp. 293–295, 2016.
- [15] S. Eguchi, “Chapter 2 - pythagoras theorem in information geometry and applications to generalized linear models,” in *Information Geometry*, ser. Handbook of Statistics, A. Plastino, A. S. R. S. Rao, and C. R. Rao, Eds. Elsevier, 2021, vol. 45, pp. 15–42. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0169716121000225>
- [16] J. Kim, J. Kwon, M. Cho, H. Lee, and J.-H. Won, “ t^3 -variational autoencoder: Learning heavy-tailed data with student’s t and power divergence,” *arXiv preprint arXiv:2312.01133*, 2023.