

Diffusion Models as Stochastic Differential Equations

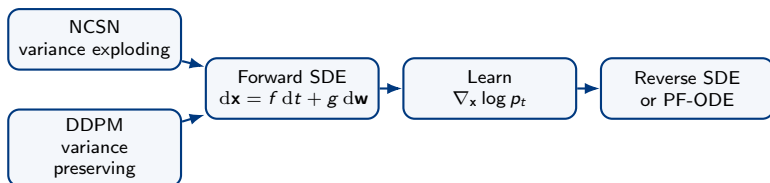
Forward Diffusion, Time Reversal, Training, and Variational Connections

Heejoon Byun

SNU, statistics dept.

July 2026

One continuous object unifies NCSN and DDPM



Central message

Continuous time separates diffusion modeling into three mathematical jobs: specify a forward perturbation, learn its marginal score, and integrate a reverse differential equation.

Table of contents

- 1 Introduction: Motivation and Basic Definitions
- 2 Forward-Time SDEs
- 3 Reverse-Time Processes for Generation
- 4 Training and Sampling
- 5 Different Instantiations and Variational Formulation
- 6 Summary

Introduction

How do DDPM and NCSN become SDEs?

NCSN and DDPM share an Euler–Maruyama skeleton

On a grid with step Δt , both forward processes have Gaussian increments.

$$\begin{aligned}\text{NCSN: } \mathbf{x}_{t+\Delta t} &= \mathbf{x}_t + \sqrt{\sigma^2(t+\Delta t) - \sigma^2(t)} \boldsymbol{\epsilon}_t \\ &\approx \mathbf{x}_t + \sqrt{\frac{d\sigma^2(t)}{dt}} \sqrt{\Delta t} \boldsymbol{\epsilon}_t,\end{aligned}$$

$$\begin{aligned}\text{DDPM: } \mathbf{x}_{t+\Delta t} &= \sqrt{1 - \beta(t)\Delta t} \mathbf{x}_t + \sqrt{\beta(t)\Delta t} \boldsymbol{\epsilon}_t \\ &\approx \mathbf{x}_t - \frac{1}{2}\beta(t)\mathbf{x}_t\Delta t + \sqrt{\beta(t)}\sqrt{\Delta t} \boldsymbol{\epsilon}_t,\end{aligned}$$

where $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, I)$.

Common local form

$$\mathbf{x}_{t+\Delta t} \approx \mathbf{x}_t + f(\mathbf{x}_t, t)\Delta t + g(t)\sqrt{\Delta t} \boldsymbol{\epsilon}_t.$$

The continuous limit records drift and local variance

Sending $\Delta t \rightarrow 0$ gives the forward SDE

$$d\mathbf{x}(t) = f(\mathbf{x}(t), t) dt + g(t) d\mathbf{w}(t), \quad \mathbf{x}(0) \sim p_{\text{data}}.$$

Model	$f(\mathbf{x}, t)$	$g(t)$
NCSN / VE	0	$\sqrt{d\sigma^2(t)/dt}$
DDPM / VP	$-\frac{1}{2}\beta(t)\mathbf{x}$	$\sqrt{\beta(t)}$

Interpretation

f controls systematic motion; g^2 is the instantaneous rate at which covariance is injected. Different pairs (f, g) create different noise paths.

Forward diffusion transports data to a simple prior

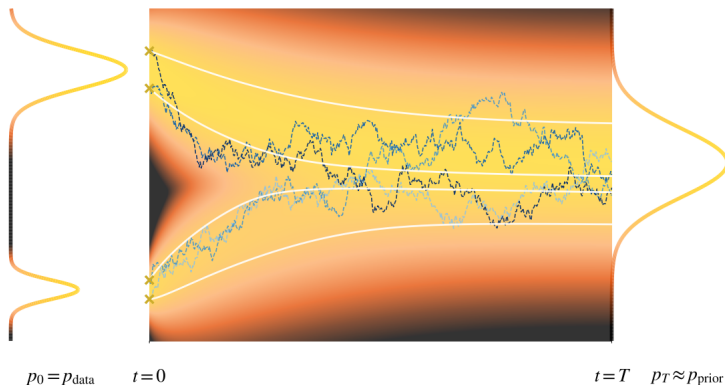


Figure 4.3. The heatmap shows p_t : stochastic forward-SDE paths (blue) and deterministic PF-ODE paths (white) share the same evolving marginals from $p_0 = p_{\text{data}}$ to $p_T \approx p_{\text{prior}}$.

A Wiener process is continuous but has rough increments

$$d\mathbf{x}(t) = f(\mathbf{x}(t), t) dt + g(t) d\mathbf{w}(t).$$

A standard D -dimensional Wiener process $\mathbf{w}(t)$ satisfies

- 1 $\mathbf{w}(0) = 0$ almost surely;
- 2 increments over disjoint intervals are independent;
- 3 $\mathbf{w}(t) - \mathbf{w}(s) \sim \mathcal{N}(0, (t - s)I_D)$ for $0 \leq s < t$;
- 4 sample paths are almost surely continuous and nowhere differentiable.

Integral meaning of an SDE

$$\mathbf{x}(t) = \mathbf{x}(0) + \int_0^t f(\mathbf{x}(s), s) ds + \int_0^t g(s) d\mathbf{w}(s).$$

The symbol $d\mathbf{w}(t)$ abbreviates an increment with covariance $dt I_D$; it is not an ordinary differential.

Itô integration is built from left-endpoint sums

For a simple process on a partition $0 = t_0 < \dots < t_n = t$,

$$g(s) = \sum_{i=0}^{n-1} g(t_i) \mathbf{1}_{[t_i, t_{i+1})}(s),$$

define

$$\int_0^t g(s) d\mathbf{w}(s) := \sum_{i=0}^{n-1} g(t_i) (\mathbf{w}(t_{i+1}) - \mathbf{w}(t_i)).$$

General Itô integrals are mean-square limits of such sums. Left-endpoint evaluation prevents the integrand from using future noise.

Forward SDE

How can an SDE interpolate data and noise laws?

Linear drift makes every perturbation kernel Gaussian

Consider

$$d\mathbf{x}(t) = f(t)\mathbf{x}(t) dt + g(t) d\mathbf{w}(t).$$

Conditional on $\mathbf{x}(0) = \mathbf{x}_0$, its perturbation kernel is

$$p_t(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; m(t), P(t)I_D),$$

with

$$\begin{aligned} m(t) &= \exp\left(\int_0^t f(u) du\right) \mathbf{x}_0, \\ P(t) &= \int_0^t \exp\left(2 \int_s^t f(u) du\right) g^2(s) ds. \end{aligned}$$

Why this matters

Training can sample $\mathbf{x}_t$ directly from $\mathbf{x}_0$; no SDE simulation is needed along the forward path.

Marginals are mixtures of perturbation kernels

At the initial time, the process has not perturbed the data:

$$m(0) = \mathbf{x}_0, \quad P(0) = 0, \quad p_0(\mathbf{x} \mid \mathbf{x}_0) = \delta_{\mathbf{x}_0}(\mathbf{x}).$$

The forward SDE induces

$$p_t(\mathbf{x}_t) = \int p_t(\mathbf{x}_t \mid \mathbf{x}_0) p_{\text{data}}(\mathbf{x}_0) d\mathbf{x}_0, \quad p_0 = p_{\text{data}}.$$

Endpoint conditions

$$p_T(\mathbf{x}_T \mid \mathbf{x}_0) \approx p_{\text{prior}}(\mathbf{x}_T) \implies p_T(\mathbf{x}_T) \approx p_{\text{prior}}(\mathbf{x}_T).$$

A useful terminal design makes the kernel forget its initial state:

$$\exp\left(\int_0^T f(u) du\right) \mathbf{x}_0 \approx 0, \quad P(T)I_D \approx \text{covariance of } p_{\text{prior}}.$$

The conditional kernel is Gaussian; the marginal p_t is generally a non-Gaussian mixture and its score is unknown.

Quadratic variation explains the identity $(dw)^2 = dt$

For one-dimensional Brownian motion on $[0, t]$, let

$$Q_{\Pi} := \sum_i (w(t_{i+1}) - w(t_i))^2$$

for a partition Π . Independent Gaussian increments give

$$\mathbb{E}[Q_{\Pi}] = \sum_i (t_{i+1} - t_i) = t,$$

$$\text{Var}(Q_{\Pi}) = 2 \sum_i (t_{i+1} - t_i)^2 \leq 2t |\Pi|.$$

Hence $Q_{\Pi} \rightarrow t$ in L^2 as the mesh $|\Pi| \rightarrow 0$.

Itô multiplication table

$$(dw)^2 = dt, \quad dt \, dw = 0, \quad (dt)^2 = 0.$$

The first identity is shorthand for convergence of quadratic variation, not pointwise algebra of infinitesimals.

Itô's formula keeps the second-order Brownian term

For a scalar Itô process x_t and a scalar function h ,

$$dh(x_t, t) = \partial_t h \, dt + \partial_x h \, dx_t + \frac{1}{2} \partial_{xx} h (dx_t)^2.$$

If $dx_t = f(x_t, t) \, dt + g(t) \, dw_t$, then $(dx_t)^2 = g^2(t) \, dt$, so

$$dh(x_t, t) = \left(\partial_t h + f \partial_x h + \frac{1}{2} g^2 \partial_{xx} h \right) dt + g \partial_x h \, dw_t.$$

This is the second-order Taylor correction induced by quadratic variation. For $\mathbf{x}_t \in \mathbb{R}^D$ and scalar h ,

$$dh(\mathbf{x}_t, t) = \left(\partial_t h + \nabla h^\top f + \frac{1}{2} g^2(t) \text{Tr}(\nabla^2_{\mathbf{x}} h) \right) dt + g(t) \nabla h^\top \, d\mathbf{w}_t.$$

An integrating factor gives the linear SDE solution

For $d\mathbf{x}_t = f(t)\mathbf{x}_t dt + g(t) d\mathbf{w}_t$, set

$$\Psi(t) := \exp\left(-\int_0^t f(s) ds\right), \quad \mathbf{y}_t := \Psi(t)\mathbf{x}_t.$$

Since Ψ is deterministic, applying Itô's formula to $h(t, \mathbf{x}) = \Psi(t)\mathbf{x}$ gives

$$\begin{aligned} d\mathbf{y}_t &= \Psi'(t)\mathbf{x}_t dt + \Psi(t) d\mathbf{x}_t \\ &= -f(t)\Psi(t)\mathbf{x}_t dt + \Psi(t)[f(t)\mathbf{x}_t dt + g(t) d\mathbf{w}_t] \\ &= \Psi(t)g(t) d\mathbf{w}_t. \end{aligned}$$

Therefore

$$\mathbf{x}_t = e^{\int_0^t f(s) ds} \mathbf{x}_0 + \int_0^t e^{\int_s^t f(u) du} g(s) d\mathbf{w}_s.$$

Itô isometry yields the kernel variance

Write

$$E(s \rightarrow t) := \exp\left(\int_s^t f(u) du\right).$$

Then

$$\mathbf{x}_t = E(0 \rightarrow t)\mathbf{x}_0 + \int_0^t E(s \rightarrow t)g(s) d\mathbf{w}_s.$$

If $\phi(s)$ is a predictable stochastic process satisfying

$$\mathbb{E} \int_0^t \|\phi(s)\|_F^2 ds < \infty,$$

then its Itô integral has mean zero. The *Itô isometry* states

$$\mathbb{E} \left\| \int_0^t \phi(s) d\mathbf{w}_s \right\|^2 = \mathbb{E} \int_0^t \|\phi(s)\|_F^2 ds.$$

Here $\phi(s) = E(s \rightarrow t)g(s)I_D$ is deterministic, so the integral is also Gaussian.

Therefore

$$\begin{aligned} \mathbb{E}[\mathbf{x}_t \mid \mathbf{x}_0] &= E(0 \rightarrow t)\mathbf{x}_0, \\ \text{Cov}(\mathbf{x}_t \mid \mathbf{x}_0) &= \left[\int_0^t E^2(s \rightarrow t)g^2(s) ds \right] I_D. \end{aligned}$$

Reverse Process

How can the same density path run backward?

Reverse time uses a new clock and Wiener process

The customary decreasing-time notation is

$$\begin{aligned} d\bar{\mathbf{x}}(t) = & \left[f(\bar{\mathbf{x}}(t), t) - g^2(t) \nabla_{\mathbf{x}} \log p_t(\bar{\mathbf{x}}(t)) \right] dt \\ & + g(t) d\bar{\mathbf{w}}(t), \quad t : T \downarrow 0. \end{aligned}$$

Let $r = T - t$ denote elapsed reverse time. If $\mathbf{w}$ is a standard forward Wiener process, define

$$\bar{\mathbf{w}}(r) := \mathbf{w}(T - r) - \mathbf{w}(T), \quad 0 \leq r \leq T.$$

For $0 \leq r_1 < r_2 \leq T$,

$$\bar{\mathbf{w}}(r_2) - \bar{\mathbf{w}}(r_1) = \mathbf{w}(T - r_2) - \mathbf{w}(T - r_1) \sim \mathcal{N}(0, (r_2 - r_1)I_D),$$

and disjoint increments are independent. Hence $\bar{\mathbf{w}}$ is a standard Wiener process when indexed by the increasing reverse clock r .

Same one-time marginals

If $\bar{\mathbf{x}}(T) \sim p_T$ and the exact score is used, then $\bar{\mathbf{x}}(t) \sim p_t$ for every $t \in [0, T]$. Thus the forward and reverse SDEs have the same marginal density at each time, although their path laws and directions of evolution differ.

Reverse diffusion reconstructs the data distribution

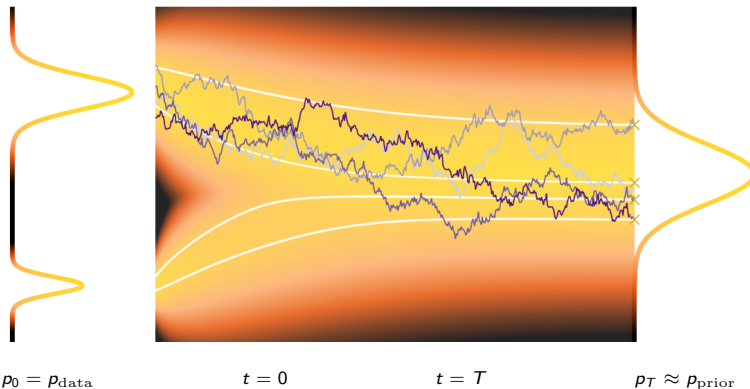


Figure 4.4. Reverse-SDE trajectories start from $p_T \approx p_{\text{prior}}$ at $t = T$ and collectively recover the bimodal data distribution $p_0 = p_{\text{data}}$ at $t = 0$.

The probability flow ODE follows the same marginals

Define

$$v(\mathbf{x}, t) := f(\mathbf{x}, t) - \frac{1}{2}g^2(t)\nabla_{\mathbf{x}} \log p_t(\mathbf{x}).$$

The probability flow ODE is

$$\frac{d\tilde{\mathbf{x}}(t)}{dt} = v(\tilde{\mathbf{x}}(t), t).$$

Same densities, different paths

At each t , the PF-ODE and forward SDE have the same marginal density p_t ; integrating the ODE backward from p_T therefore generates p_0 .

- The PF-ODE is deterministic and invertible under regularity.
- The factor $1/2$ is essential: the ODE drift replaces the density spreading that diffusion produced.
- The reverse SDE uses the full factor g^2 because it still diffuses.

Fokker-Planck governs the marginal density

If

$$d\mathbf{x}_t = f(\mathbf{x}_t, t) dt + g(t) d\mathbf{w}_t,$$

then its marginal density satisfies

Fokker-Planck equation

$$\partial_t p_t(\mathbf{x}) = -\nabla_{\mathbf{x}} \cdot [f(\mathbf{x}, t) p_t(\mathbf{x})] + \frac{1}{2} g^2(t) \Delta_{\mathbf{x}} p_t(\mathbf{x}), \quad p_0 = p_{\text{data}}.$$

The two terms have distinct roles:

- $-\nabla \cdot (f p_t)$ transports probability under the drift;
- $\frac{1}{2} g^2 \Delta p_t$ spreads probability under Brownian noise.

Continuity equation: the zero-diffusion case

For the deterministic ODE $d\mathbf{x}/dt = v(\mathbf{x}, t)$, set $g = 0$ and $f = v$:

$$\partial_t p_t = -\nabla \cdot (v p_t).$$

This transport PDE is called the *continuity equation*.

Theorem 4.2.1: Fokker–Planck aligns all three processes

Theorem 4.2.1 (marginal alignment)

Let the forward SDE start from $p_0 = p_{\text{data}}$ and have marginal densities p_t . Then

$$\partial_t p_t = -\nabla \cdot (f p_t) + \frac{1}{2} g^2 \Delta p_t = -\nabla \cdot (v p_t), \quad v = f - \frac{1}{2} g^2 \nabla \log p_t.$$

Consequently:

- 1 the PF-ODE $d\tilde{\mathbf{x}}/dt = v(\tilde{\mathbf{x}}, t)$ has $\tilde{\mathbf{x}}(t) \sim p_t$ when run forward from p_0 , and also when run backward from p_T ;
- 2 the reverse-time SDE

$$d\bar{\mathbf{x}} = [f - g^2 \nabla \log p_t] dt + g d\bar{\mathbf{w}}$$

started from p_T has the time-reversed marginal family.

SDE, Fokker–Planck, and marginal alignment

An SDE specifies random paths; its Fokker–Planck equation is the deterministic conservation law for their density. If two processes solve the same density PDE from the same endpoint density, uniqueness forces the same p_t at every time—even though their individual paths differ.

Smoothness and uniqueness of the associated PDEs are assumed.

Fokker-Planck proof I: weak form

Take a spatial test function $\phi \in C_c^2(\mathbb{R}^D)$. For

$$d\mathbf{x}_t = f(\mathbf{x}_t, t) dt + g(t) d\mathbf{w}_t,$$

Itô's formula gives

$$d\phi(\mathbf{x}_t) = \left[\nabla \phi(\mathbf{x}_t)^\top f(\mathbf{x}_t, t) + \frac{1}{2} g^2(t) \Delta \phi(\mathbf{x}_t) \right] dt + g(t) \nabla \phi(\mathbf{x}_t)^\top d\mathbf{w}_t.$$

Integrate from t to $t + h$ and take expectations. The Itô integral has mean zero, so division by h followed by $h \downarrow 0$ yields

$$\frac{d}{dt} \mathbb{E}[\phi(\mathbf{x}_t)] = \mathbb{E} \left[\nabla \phi(\mathbf{x}_t)^\top f(\mathbf{x}_t, t) + \frac{1}{2} g^2(t) \Delta \phi(\mathbf{x}_t) \right].$$

Writing expectations with the density p_t gives the weak identity

$$\int \phi(\mathbf{x}) \partial_t p_t(\mathbf{x}) d\mathbf{x} = \int \left[\nabla \phi(\mathbf{x})^\top f(\mathbf{x}, t) + \frac{1}{2} g^2(t) \Delta \phi(\mathbf{x}) \right] p_t(\mathbf{x}) d\mathbf{x}.$$

Compact support justifies the next spatial integrations by parts.

Fokker–Planck proof II: integration by parts

Because ϕ has compact support,

$$\int \nabla \phi^\top f p_t \, d\mathbf{x} = - \int \phi \nabla \cdot (f p_t) \, d\mathbf{x},$$

and two integrations by parts give

$$\int \Delta \phi p_t \, d\mathbf{x} = \int \phi \Delta p_t \, d\mathbf{x}.$$

Substitution into the weak identity gives

$$\int \phi \left[\partial_t p_t + \nabla \cdot (f p_t) - \frac{1}{2} g^2(t) \Delta p_t \right] d\mathbf{x} = 0.$$

Since this holds in particular for every $\phi \in C_c^\infty(\mathbb{R}^D)$, the fundamental lemma implies, pointwise whenever the displayed derivatives are continuous,

$$\partial_t p_t(\mathbf{x}) = -\nabla_{\mathbf{x}} \cdot (f(\mathbf{x}, t) p_t(\mathbf{x})) + \frac{1}{2} g^2(t) \Delta_{\mathbf{x}} p_t(\mathbf{x}).$$

This is the Fokker–Planck equation. □

Training

How can we learn the score and sample from it?

Denoising score matching replaces an intractable target

The ideal score-matching objective uses the unknown marginal score:

$$\mathcal{L}_{\text{SM}}(\phi) = \frac{1}{2} \mathbb{E}_{t \sim p_{\text{time}}, \mathbf{x}_t \sim p_t} \left[\omega(t) \|s_\phi(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)\|^2 \right].$$

For linear SDEs, the conditional score of the Gaussian kernel is known:

$$\mathcal{L}_{\text{DSM}}(\phi) = \frac{1}{2} \mathbb{E}_{t, \mathbf{x}_0, \mathbf{x}_t | \mathbf{x}_0} \left[\omega(t) \|s_\phi(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t | \mathbf{x}_0)\|^2 \right].$$

Statistical view

DSM is a conditional least-squares regression whose noisy target has the desired marginal score as its posterior mean.

The marginal score is a posterior-weighted average

Proposition 4.3.1 (minimizer of DSM)

For almost every $\mathbf{x}_t \sim p_t$ and $t \in [0, T]$,

$$s^*(\mathbf{x}_t, t) = \mathbb{E}_{\mathbf{x}_0 \sim p(\mathbf{x}_0 | \mathbf{x}_t)} [\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t | \mathbf{x}_0)] = \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t).$$

The “weights” are posterior responsibilities: clean samples more likely to have produced the observed $\mathbf{x}_t$ contribute more.

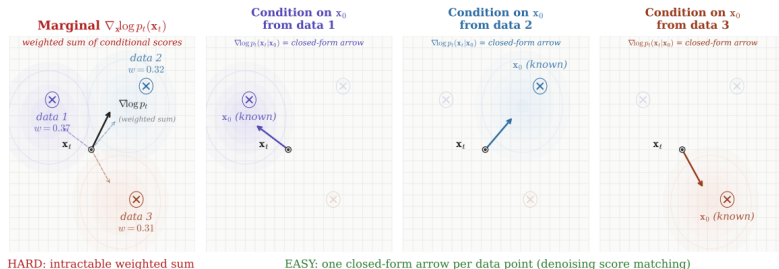
For the empirical data distribution $\hat{p}_{\text{data}} = \frac{1}{n} \sum_{i=1}^n \delta_{\mathbf{x}_0^{(i)}}$,

$$\begin{aligned} \nabla_{\mathbf{x}_t} \log \hat{p}_t(\mathbf{x}_t) &= \sum_{i=1}^n w_i(\mathbf{x}_t, t) \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t | \mathbf{x}_0^{(i)}), \\ w_i(\mathbf{x}_t, t) &= \frac{p_t(\mathbf{x}_t | \mathbf{x}_0^{(i)})}{\sum_{j=1}^n p_t(\mathbf{x}_t | \mathbf{x}_0^{(j)})}, \quad \sum_{i=1}^n w_i(\mathbf{x}_t, t) = 1. \end{aligned}$$

Why DSM is tractable

One sampled clean point provides one closed-form conditional-score target. Squared-error regression recovers its posterior mean: the marginal score.

Figure 4.6: conditional scores average to the marginal score



Each conditional density supplies an easy analytic score field. At a noisy point $\mathbf{x}_t$, the marginal score is their posterior-weighted vector average. DSM learns this average without evaluating the mixture density.

Sampling discretizes either the reverse SDE or the PF-ODE

With step $\Delta t > 0$ from t to $t - \Delta t$ and $s_\phi \approx \nabla \log p_t$:

Euler–Maruyama: empirical reverse SDE

$$\mathbf{x}_{t-\Delta t} = \mathbf{x}_t - [f(\mathbf{x}_t, t) - g^2(t)s_\phi(\mathbf{x}_t, t)]\Delta t + g(t)\sqrt{\Delta t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I).$$

Euler: empirical PF-ODE

$$\mathbf{x}_{t-\Delta t} = \mathbf{x}_t - \left[f(\mathbf{x}_t, t) - \frac{1}{2}g^2(t)s_\phi(\mathbf{x}_t, t) \right] \Delta t.$$

	Reverse SDE	PF-ODE
Path	stochastic	deterministic
Marginal target	p_t	p_t
Numerical error	SDE discretization	ODE integration

Instantiations

How can SDEs realize prescribed probability kernels?

VE SDE is the continuous limit of the NCSN noise ladder

Set $f = 0$ and $g(t) = \sqrt{d\sigma^2(t)/dt}$:

$$d\mathbf{x}(t) = \sqrt{\frac{d\sigma^2(t)}{dt}} d\mathbf{w}(t).$$

The kernel and its score are

$$p_t(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \mathbf{x}_0, [\sigma^2(t) - \sigma^2(0)]I_D),$$

$$\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t | \mathbf{x}_0) = -\frac{\mathbf{x}_t - \mathbf{x}_0}{\sigma^2(t) - \sigma^2(0)}.$$

Why “variance exploding”?

The signal is not contracted: $\mathbb{E}[\mathbf{x}_t | \mathbf{x}_0] = \mathbf{x}_0$. Meanwhile,

$$\text{Cov}(\mathbf{x}_t | \mathbf{x}_0) = [\sigma^2(t) - \sigma^2(0)]I_D, \quad \text{Cov}(\mathbf{x}_t) = \text{Cov}(\mathbf{x}_0) + [\sigma^2(t) - \sigma^2(0)]I_D.$$

Increasing $\sigma(t)$ adds variance without attenuating the signal, so the total marginal variance grows with time.

NCSN schedule

A geometric $\sigma(t) = \sigma_{\min}(\sigma_{\max}/\sigma_{\min})^t$ spans many noise scales; for large $\sigma(T)$, $p_T \approx \mathcal{N}(0, \sigma^2(T)I_D)$ after centering data.

VP SDE is the continuous-time DDPM limit

Let $\beta(t) \geq 0$ and

$$d\mathbf{x}(t) = -\frac{1}{2}\beta(t)\mathbf{x}(t)dt + \sqrt{\beta(t)}d\mathbf{w}(t).$$

With $B(t) = \int_0^t \beta(s) ds$,

$$p_t(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t; e^{-B(t)/2}\mathbf{x}_0, [1 - e^{-B(t)}]I_D\right),$$

$$\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t | \mathbf{x}_0) = -\frac{\mathbf{x}_t - e^{-B(t)/2}\mathbf{x}_0}{1 - e^{-B(t)}}.$$

Why “variance preserving”?

The conditional noise variance $1 - e^{-B(t)}$ grows, but the drift contracts the signal by $e^{-B(t)/2}$. Hence

$$\text{Cov}(\mathbf{x}_t) = e^{-B(t)} \text{Cov}(\mathbf{x}_0) + [1 - e^{-B(t)}]I_D.$$

$$\text{Cov}(\mathbf{x}_0) = I_D \implies \text{Cov}(\mathbf{x}_t) = I_D \text{ for every } t.$$

The preserved quantity is the *marginal variance*, not the conditional noise variance.

DDPM schedule

A common choice is $\beta(t) = \beta_{\min} + t(\beta_{\max} - \beta_{\min})$. As $B(T)$ grows, $p_T \approx \mathcal{N}(0, I_D)$.

Which samples faster: VE or VP?

Continuous time: equivalent

For an independent Gaussian affine path,

$$\tilde{\mathbf{x}} := \frac{\mathbf{x}}{\alpha_t}, \quad \rho := \frac{\sigma_t}{\alpha_t}, \quad \tilde{\mathbf{x}}_\rho = \mathbf{x}_0 + \rho \epsilon,$$

and a consistent score transformation give

$$\frac{d\tilde{\mathbf{x}}}{d\rho} = -\rho \nabla_{\tilde{\mathbf{x}}} \log p_\rho(\tilde{\mathbf{x}}).$$

Exact statement

With the exact score, VE and VP trajectories satisfy $\mathbf{x}_t = \alpha_t \tilde{\mathbf{x}}_{\rho(t)}$. Matched ρ -nodes and the same normalized solver give identical iterates.

Finite NFE: not invariant

For an order- p method,

$$\|\tau_i\| = O\left(h_i^{p+1} \sup_{[t_i, t_{i+1}]} \|\mathbf{x}^{(p+1)}(t)\|\right).$$

- uniform t -grids induce different ρ -grids;
- raw-coordinate stiffness and learned score errors differ.

Why VP often wins by default

The bounded scale $\alpha_t^2 + \sigma_t^2 \approx 1$ improves conditioning. Standard VP grids, parameterizations, and samplers are jointly tuned, whereas naive VE spans a larger noise range. Thus fixed-NFE legacy comparisons often favor VP.

Direct answer

Exact normalized dynamics are equivalent, but fixed-NFE implementations are not. VP is often safer by default; a tuned VE pipeline can narrow the practical gap.

Song et al. (2021); Karras et al. (2022); Ghanem and Bzdok (2024); Yang et al. (2024); Cui et al. (2025).

Affine kernels unify VE, VP, and flow matching

Specify the accumulated perturbation directly:

$$p_t(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \alpha_t \mathbf{x}_0, \sigma_t^2 I_D),$$

$$\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \epsilon.$$

Typically $\alpha_0 = 1$, $\sigma_0 = 0$, and $\text{SNR}(t) := \alpha_t^2 / \sigma_t^2$ decreases with t .

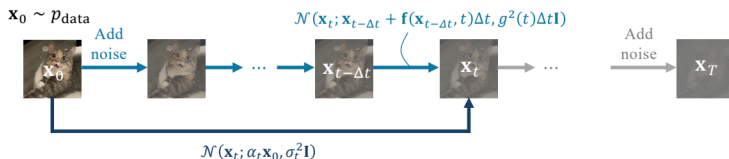


Figure 4.8. Incremental noise injection by a continuous-time SDE and direct sampling from the accumulated affine kernel are equivalent.

Mean and variance satisfy closed ODEs

For

$$d\mathbf{x}_t = f(t)\mathbf{x}_t dt + g(t) d\mathbf{w}_t,$$

define

$$m(t) := \mathbb{E}[\mathbf{x}_t \mid \mathbf{x}_0], \quad P(t)I_D := \text{Cov}(\mathbf{x}_t \mid \mathbf{x}_0).$$

Moment ODEs

$$m'(t) = f(t)m(t), \quad P'(t) = 2f(t)P(t) + g^2(t),$$

with $m(0) = \mathbf{x}_0$ and $P(0) = 0$.

For the mean, take conditional expectations in the SDE:

$$dm(t) = f(t)\mathbb{E}[\mathbf{x}_t \mid \mathbf{x}_0] dt,$$

because the Itô integral has conditional mean zero.

Proof: Itô's formula gives the variance ODE

Let $\tilde{\mathbf{x}}_t := \mathbf{x}_t - m(t)$. The mean equation implies

$$d\tilde{\mathbf{x}}_t = f(t)\tilde{\mathbf{x}}_t dt + g(t) d\mathbf{w}_t.$$

Apply Itô's formula componentwise to $h_{ij}(\tilde{\mathbf{x}}) = \tilde{x}_i \tilde{x}_j$:

$$\begin{aligned} d(\tilde{\mathbf{x}}_t \tilde{\mathbf{x}}_t^\top) &= d\tilde{\mathbf{x}}_t \tilde{\mathbf{x}}_t^\top + \tilde{\mathbf{x}}_t d\tilde{\mathbf{x}}_t^\top + d\tilde{\mathbf{x}}_t d\tilde{\mathbf{x}}_t^\top \\ &= \left[2f(t)\tilde{\mathbf{x}}_t \tilde{\mathbf{x}}_t^\top + g^2(t)I_D \right] dt + \text{zero-mean Itô terms.} \end{aligned}$$

Taking conditional expectations yields

$$\frac{d}{dt}(P(t)I_D) = 2f(t)P(t)I_D + g^2(t)I_D.$$

Hence $P'(t) = 2f(t)P(t) + g^2(t)$. □

Lemma 4.5.1: the affine kernel determines the linear SDE

Lemma 4.5.1 (forward kernel $\iff$ linear SDE)

Define $\lambda_t := \log(\alpha_t/\sigma_t)$ for $t \in (0, T]$. The SDE

$$d\mathbf{x}(t) = f(t)\mathbf{x}(t) dt + g(t) d\mathbf{w}(t)$$

has transition $\mathcal{N}(\mathbf{x}_t; \alpha_t \mathbf{x}_0, \sigma_t^2 I_D)$ when

$$f(t) = \frac{d}{dt} \log \alpha_t,$$

$$g^2(t) = \frac{d\sigma_t^2}{dt} - 2 \frac{d}{dt} \log \alpha_t \sigma_t^2 = -2\sigma_t^2 \frac{d\lambda_t}{dt}.$$

Conversely, a linear SDE with these Gaussian transitions satisfies the same coefficient identities.

Since $\lambda_t = \frac{1}{2} \log \text{SNR}(t)$, the requirement $g^2(t) \geq 0$ is equivalent to nonincreasing log-SNR.

α and σ give a pathwise solution

Recall the integrating-factor solution in terms of drift and diffusion:

$$\mathbf{x}_t = e^{\int_0^t f(r) dr} \mathbf{x}_0 + \int_0^t e^{\int_u^t f(r) dr} g(u) d\mathbf{w}_u.$$

Lemma 4.5.1 gives

$$f(u) = \frac{d}{du} \log \alpha_u, \quad g^2(u) = \alpha_u^2 \frac{d}{du} \left(\frac{\sigma_u^2}{\alpha_u^2} \right).$$

Substitution eliminates f and g :

$$\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \alpha_t \int_0^t \sqrt{\frac{d}{du} \left(\frac{\sigma_u^2}{\alpha_u^2} \right)} d\mathbf{w}_u.$$

Here $\alpha_0 = 1$ and $\sigma_0 = 0$. Itô isometry verifies

$$\text{Cov}(\mathbf{x}_t \mid \mathbf{x}_0) = \alpha_t^2 \int_0^t \frac{d}{du} \left(\frac{\sigma_u^2}{\alpha_u^2} \right) du I_D = \sigma_t^2 I_D.$$

The Markov property gives intermediate transitions

For $0 \leq t < s \leq T$, split the stochastic integral at t :

$$\begin{aligned} \mathbf{x}_s &= \alpha_s \left[\mathbf{x}_0 + \int_0^t \sqrt{\frac{d}{du} \left(\frac{\sigma_u^2}{\alpha_u^2} \right)} d\mathbf{w}_u \right] + \alpha_s \int_t^s \sqrt{\frac{d}{du} \left(\frac{\sigma_u^2}{\alpha_u^2} \right)} d\mathbf{w}_u \\ &= \frac{\alpha_s}{\alpha_t} \mathbf{x}_t + \alpha_s \int_t^s \sqrt{\frac{d}{du} \left(\frac{\sigma_u^2}{\alpha_u^2} \right)} d\mathbf{w}_u. \end{aligned}$$

The bracket in the first line equals $\mathbf{x}_t/\alpha_t$. The remaining integral uses only Brownian increments after t and is independent of the information available at time t . Thus the SDE solution is Markov: conditional on $\mathbf{x}_t$, the earlier path gives no further information about $\mathbf{x}_s$.

$$\alpha_{s|t} := \frac{\alpha_s}{\alpha_t}, \quad \sigma_{s|t}^2 := \alpha_s^2 \int_t^s \frac{d}{du} \left(\frac{\sigma_u^2}{\alpha_u^2} \right) du = \sigma_s^2 - \alpha_{s|t}^2 \sigma_t^2.$$

Markov transition

$$p(\mathbf{x}_s \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_s; \alpha_{s|t} \mathbf{x}_t, \sigma_{s|t}^2 I_D).$$

Lemma 4.5.2 gives the reverse conditional kernel

Lemma 4.5.2 (reverse conditional transitions)

Let $0 \leq t < s \leq T$ and condition on a clean value $\mathbf{x}$. Then

$$p(\mathbf{x}_t \mid \mathbf{x}_s, \mathbf{x}) = \mathcal{N}(\mathbf{x}_t; \mu(\mathbf{x}_s, \mathbf{x}; s, t), \sigma^2(s, t)I_D),$$

where

$$\mu(\mathbf{x}_s, \mathbf{x}; s, t) = \frac{\alpha_{s|t}\sigma_t^2}{\sigma_s^2}\mathbf{x}_s + \frac{\alpha_t\sigma_{s|t}^2}{\sigma_s^2}\mathbf{x},$$

$$\sigma^2(s, t) = \frac{\sigma_{s|t}^2\sigma_t^2}{\sigma_s^2},$$

$$\alpha_{s|t} = \frac{\alpha_s}{\alpha_t}, \quad \sigma_{s|t}^2 = \sigma_s^2 - \alpha_{s|t}^2\sigma_t^2.$$

No variance-preserving constraint $\alpha_t^2 + \sigma_t^2 = 1$ is required.

Proof: derive the joint conditional Gaussian

Conditional on the clean value $\mathbf{x}$,

$$\mathbf{x}_t \sim \mathcal{N}(\alpha_t \mathbf{x}, \sigma_t^2 I_D), \quad \mathbf{x}_s | \mathbf{x}_t \sim \mathcal{N}(\alpha_{s|t} \mathbf{x}_t, \sigma_{s|t}^2 I_D).$$

The second relation follows from the Markov property. Its innovation is independent of $\mathbf{x}_t$, so

$$\text{Cov}(\mathbf{x}_t, \mathbf{x}_s | \mathbf{x}) = \alpha_{s|t} \sigma_t^2 I_D.$$

Therefore

$$\begin{pmatrix} \mathbf{x}_t \\ \mathbf{x}_s \end{pmatrix} | \mathbf{x} \sim \mathcal{N} \left(\begin{pmatrix} \alpha_t \mathbf{x} \\ \alpha_s \mathbf{x} \end{pmatrix}, \begin{pmatrix} \sigma_t^2 I_D & \alpha_{s|t} \sigma_t^2 I_D \\ \alpha_{s|t} \sigma_t^2 I_D & \sigma_s^2 I_D \end{pmatrix} \right).$$

Variational reversal learns a clean-signal predictor

Lemma 4.5.2 suggests the $\mathbf{x}$ -prediction parameterization

$$p_{\phi}(\mathbf{x}_t \mid \mathbf{x}_s) := \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_{\phi}(\mathbf{x}_s, s, t), \sigma^2(s, t)I_D),$$

with

$$\boldsymbol{\mu}_{\phi}(\mathbf{x}_s, s, t) = \frac{\alpha_{s|t}\sigma_t^2}{\sigma_s^2}\mathbf{x}_s + \frac{\alpha_t\sigma_{s|t}^2}{\sigma_s^2}\mathbf{x}_{\phi}(\mathbf{x}_s, s).$$

Equivalently, predict noise through

$$\mathbf{x}_s = \alpha_s \mathbf{x}_{\phi}(\mathbf{x}_s, s) + \sigma_s \boldsymbol{\epsilon}_{\phi}(\mathbf{x}_s, s).$$

Modeling choice

The reverse variance is fixed to the exact conditional variance; learning is concentrated in a prediction of the clean signal, equivalently the noise or score.

VDM uses the forward process as a variational encoder

On $\Pi_\varepsilon = \{\varepsilon = t_0 < \dots < t_N = T\}$, take the forward chain

$$q(\mathbf{x}_{t_0:t_N} \mid \mathbf{x}_0) = q(\mathbf{x}_{t_0} \mid \mathbf{x}_0) \prod_{i=0}^{N-1} q(\mathbf{x}_{t_{i+1}} \mid \mathbf{x}_{t_i}),$$

as the variational encoder, and define the generative reverse model by

$$p_\phi(\mathbf{x}_0, \mathbf{x}_{t_0:t_N}) = p_{\text{prior}}(\mathbf{x}_T) p_\phi(\mathbf{x}_0 \mid \mathbf{x}_\varepsilon) \prod_{i=0}^{N-1} p_\phi(\mathbf{x}_{t_i} \mid \mathbf{x}_{t_{i+1}}).$$

Known encoder

Forward kernels add noise from $\mathbf{x}_0$ to $\mathbf{x}_T$. They are analytically available and require no learned inference network.

Learned decoder

Reverse kernels denoise from $\mathbf{x}_T$ to $\mathbf{x}_\varepsilon$, followed by $p_\phi(\mathbf{x}_0 \mid \mathbf{x}_\varepsilon)$.

Why this formulation is natural

The forward noising process already defines a hierarchy of latent variables. Learning its probabilistic reversal gives both a sampler and a latent-variable likelihood objective.

The VDM ELBO separates three modeling tasks

For the encoder and reverse model on Π_ε , the negative ELBO is

$$\begin{aligned}
 -\text{ELBO}_{\Pi_\varepsilon}(\mathbf{x}_0) = & \underbrace{-\mathbb{E}_q \log p_\phi(\mathbf{x}_0 \mid \mathbf{x}_\varepsilon)}_{\text{reconstruction}} \\
 & + \underbrace{D_{\text{KL}}(q(\mathbf{x}_T \mid \mathbf{x}_0) \parallel p_{\text{prior}}(\mathbf{x}_T))}_{\text{prior matching}} \\
 & + \underbrace{\sum_{i=0}^{N-1} \mathbb{E}_q D_{\text{KL}}(q(\mathbf{x}_{t_i} \mid \mathbf{x}_{t_{i+1}}, \mathbf{x}_0) \parallel p_\phi(\mathbf{x}_{t_i} \mid \mathbf{x}_{t_{i+1}}))}_{\mathcal{L}_{\text{VDM}}^{\Pi_\varepsilon}(\mathbf{x}_0)}.
 \end{aligned}$$

Interpretation

Reconstruction models the final small-noise step; prior matching makes $\mathbf{x}_T$ simple; transition KLs train the reverse dynamics at every intermediate noise level.

Each transition KL is a discrete DSM term

For $s = t_{i+1}$ and $t = t_i$, Lemma 4.5.2 and the matched reverse covariance give

$$\begin{aligned} D_{\text{KL}}(q(\mathbf{x}_{t_i} \mid \mathbf{x}_{t_{i+1}}, \mathbf{x}_0) \parallel p_\phi(\mathbf{x}_{t_i} \mid \mathbf{x}_{t_{i+1}})) \\ = \frac{1}{2} [\text{SNR}(t_i) - \text{SNR}(t_{i+1})] \|\mathbf{x}_0 - \mathbf{x}_\phi(\mathbf{x}_{t_{i+1}}, t_{i+1})\|^2. \end{aligned}$$

The induced score and the conditional score from slide 28 satisfy

$$\underbrace{s_\phi(\mathbf{x}_s, s)}_{(\alpha_s \mathbf{x}_\phi(\mathbf{x}_s, s) - \mathbf{x}_s) / \sigma_s^2} - \underbrace{\nabla_{\mathbf{x}_s} \log q(\mathbf{x}_s \mid \mathbf{x}_0)}_{(\alpha_s \mathbf{x}_0 - \mathbf{x}_s) / \sigma_s^2} = \frac{\alpha_s}{\sigma_s^2} [\mathbf{x}_\phi(\mathbf{x}_s, s) - \mathbf{x}_0].$$

Thus the two objectives are *equivalent up to an overall positive scale, with the corresponding discrete time weights*. With both score fields evaluated at $(\mathbf{x}_{t_{i+1}}, t_{i+1})$,

$$\mathcal{L}_{\text{VDM}}^{\Pi_\varepsilon} = \frac{1}{2} \sum_{i=0}^{N-1} \omega_i \mathbb{E}_q \left\| s_\phi - \nabla_{\mathbf{x}_{t_{i+1}}} \log q(\mathbf{x}_{t_{i+1}} \mid \mathbf{x}_0) \right\|^2,$$

where $\omega_i := [\text{SNR}(t_i) - \text{SNR}(t_{i+1})] \sigma_{t_{i+1}}^4 / \alpha_{t_{i+1}}^2 > 0$.

Connection to the Score SDE narrative

DSM (slide 28) chooses ω ; VDM derives ω_i from transition KLs. Both learn the score used by discrete reverse kernels, the reverse SDE, and the PF-ODE.

Summary

What is the Score SDE framework's essential logic?

One probability path unifies diffusion modeling

$$(\alpha_t, \sigma_t) \iff (f(t), g(t)) \implies p_t \xrightarrow{\text{learn } \nabla \log p_t} \{\text{reverse SDE, PF-ODE}\}.$$

- 1 **Design the path.** Affine kernels specify simulation-free noising; the corresponding SDE realizes the same kernels incrementally.
- 2 **Learn the reversal.** Conditioning turns unknown marginal scores and reverse transitions into tractable denoising regressions; DSM and VDM differ mainly in weighting and probabilistic interpretation.
- 3 **Choose the sampler.** The reverse SDE and PF-ODE follow different trajectories but share the same marginals through Fokker–Planck.

Takeaway

NCSN, DDPM, Score SDE, and VDM are different parameterizations, objectives, and solvers for one time-indexed probability path.

Can different kernels share the same endpoints?

Many Markov dynamics can have the same endpoint marginals

$$p_t(\mathbf{y}) = \int Q_{s,t}(\mathbf{y} \mid \mathbf{x}) p_s(\mathbf{x}) d\mathbf{x}, \quad (p_0, p_T) \not\Rightarrow \{Q_{s,t} : 0 \leq s < t \leq T\}.$$

$$\mathbb{P}^* = \arg \min_{\mathbb{P}: \mathbb{P}_0 = p_{\text{data}}, \mathbb{P}_T = p_{\text{prior}}} \text{KL}(\mathbb{P} \parallel \mathbb{R}).$$

A Schrödinger bridge preserves both endpoints exactly; changing the reference path law $\mathbb{R}$ changes the intermediate kernels.

A geometry-aware kernel: Riemannian Langevin diffusion

$$d\mathbf{x}_t = -\frac{1}{2} \nabla_{\mathcal{M}} U(\mathbf{x}_t) dt + d\mathbf{B}_t^{\mathcal{M}},$$

$$Q_{s,t}^{\mathcal{M}}(\mathbf{y} \mid \mathbf{x}) := \frac{\Pr(\mathbf{x}_t \in d\mathbf{y} \mid \mathbf{x}_s = \mathbf{x})}{d\text{Vol}_{\mathcal{M}}(\mathbf{y})}.$$

$$p_{\text{ref}}(\mathbf{y}) \propto e^{-U(\mathbf{y})} \quad \text{with respect to } \text{Vol}_{\mathcal{M}}.$$

This kernel stays on $\mathcal{M}$ and follows its metric and geodesics, rather than adding ambient Euclidean Gaussian noise.

Terminal law. On compact $\mathcal{M}$, $U = 0$ gives the heat kernel and uniform prior. For a smooth positive density π , $U = -\log \pi$ makes π invariant. Exact $p_T = p_{\text{prior}}$ at finite T requires endpoint control, e.g. a Schrödinger bridge with this diffusion as $\mathbb{R}$.

De Bortoli et al., *Diffusion Schrödinger Bridge*, NeurIPS 2021; De Bortoli et al., *Riemannian SGM*, NeurIPS 2022. The transition kernel is generally not closed form; RSGM uses geodesic random walks.

What conditions must α_t and σ_t satisfy?

Set

$$\lambda(t) := \log \frac{\alpha_t}{\sigma_t} = \frac{1}{2} \log \text{SNR}(t), \quad f(t) = \frac{\alpha'_t}{\alpha_t}, \quad g^2(t) = (\sigma_t^2)' - 2f(t)\sigma_t^2 = -2\sigma_t^2\lambda'(t).$$

Valid Markov kernels

With $\alpha_0 = 1, \sigma_0 = 0$ and $\alpha_t > 0$ for finite t , require

$$\sigma_{t|s}^2 := \sigma_t^2 - \frac{\alpha_t^2}{\alpha_s^2} \sigma_s^2 \geq 0$$

for every $0 \leq s < t < \infty$. For locally absolutely continuous schedules, equivalently: λ is nonincreasing and $q(t) = g^2(t) \geq 0$ almost everywhere.

Only local regularity is needed

If $\alpha, \sigma^2 \in AC_{\text{loc}}([0, \infty))$ and $\alpha_t > 0$, then, for every finite t ,

$$f \in L^1(0, t), \quad g \in L^2(0, t).$$

Thus the linear SDE is well posed on $[0, t]$; global $f \in L^1(0, \infty)$ is not required.

Exact Gaussian endpoint

For the desired prior $\mathcal{N}(0, \tau^2 I_D)$, require, as $t \rightarrow \infty$,

$$\alpha_t \rightarrow 0, \quad \sigma_t^2 \rightarrow \tau^2.$$

Then

$$p_t(\cdot \mid \mathbf{x}_0) \Rightarrow \mathcal{N}(0, \tau^2 I_D),$$

while

$$\int_0^t f(s) ds = \log \alpha_t \longrightarrow -\infty.$$

Hence $f \notin L^1(0, \infty)$; this is compatible with local well-posedness.

References

- *Diffusion Models Today: Score SDE Framework*. Chapter 4 and Appendices A.2, C.1, D.2. arXiv:2510.21890v2, 2025.
- Song, Sohl-Dickstein, Kingma, Kumar, Ermon, and Poole. *Score-Based Generative Modeling through Stochastic Differential Equations*. ICLR, 2021.
- Song and Ermon. *Generative Modeling by Estimating Gradients of the Data Distribution*. NeurIPS, 2019.
- Ho, Jain, and Abbeel. *Denoising Diffusion Probabilistic Models*. NeurIPS, 2020.
- Kingma, Salimans, Poole, and Ho. *Variational Diffusion Models*. NeurIPS, 2021.
- Karras, Aittala, Aila, and Laine. *Elucidating the Design Space of Diffusion-Based Generative Models*. NeurIPS, 2022.
- Ghanem and Bzdok. *The Uncanny Valley: A Comprehensive Analysis of Diffusion Models*. arXiv:2402.13369, 2024.
- Lu, Zhou, Bao, Chen, Li, and Zhu. *DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps*. NeurIPS, 2022.
- De Bortoli, Thornton, Heng, and Doucet. *Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling*. NeurIPS, 2021.
- De Bortoli, Mathieu, Hutchinson, Thornton, Teh, and Doucet. *Riemannian Score-Based Generative Modelling*. arXiv:2202.02763, 2022.
- Yang, Wang, Jiang, and Li. *Leveraging Drift to Improve Sample Complexity of Variance Exploding Diffusion Models*. NeurIPS, 2024.
- Cui, Zhang, Bao, and Liao. *Elucidating the Solution Space of Extended Reverse-Time SDE for Diffusion Models*. WACV, 2025.

Supplement: PF-ODE exact model log-likelihood

With the learned score, define the model velocity

$$v_\phi(\mathbf{x}, t) := f(\mathbf{x}, t) - \frac{1}{2}g^2(t)s_\phi(\mathbf{x}, t), \quad \frac{d\mathbf{x}(t)}{dt} = v_\phi(\mathbf{x}(t), t).$$

The instantaneous change-of-variables formula gives

$$\frac{d}{dt} \log p_\phi^{\text{ODE}}(\mathbf{x}(t); t) = -\nabla_{\mathbf{x}} \cdot v_\phi(\mathbf{x}(t), t).$$

For $\mathbf{x}(0) = \mathbf{x}_0$, integrate

$$\frac{d}{dt} \begin{bmatrix} \mathbf{x}(t) \\ \delta(t) \end{bmatrix} = \begin{bmatrix} v_\phi(\mathbf{x}(t), t) \\ \nabla_{\mathbf{x}} \cdot v_\phi(\mathbf{x}(t), t) \end{bmatrix}, \quad \begin{bmatrix} \mathbf{x}(0) \\ \delta(0) \end{bmatrix} = \begin{bmatrix} \mathbf{x}_0 \\ 0 \end{bmatrix}.$$

Exact model change-of-variables likelihood

$$\log p_\phi^{\text{ODE}}(\mathbf{x}_0; 0) = \log p_{\text{prior}}(\mathbf{x}(T)) + \delta(T).$$

This identity is exact for the ODE-defined model; numerical ODE integration and divergence estimation may introduce approximation error. The proof follows from the continuity equation

$$\partial_t p_t = -\nabla_{\mathbf{x}} \cdot (v_\phi p_t).$$

Supplement: refining time gives continuous VDM

Fix $\varepsilon > 0$ and write

$$R_s(\mathbf{x}_0, \epsilon) := \|\mathbf{x}_0 - \mathbf{x}_\phi(\alpha_s \mathbf{x}_0 + \sigma_s \epsilon, s)\|^2.$$

Assume $\text{SNR} \in C^1([\varepsilon, T])$, R_s is continuous in s almost surely, and

$$\sup_{s \in [\varepsilon, T]} R_s(\mathbf{x}_0, \epsilon) \leq Z(\mathbf{x}_0, \epsilon), \quad \mathbb{E}Z < \infty.$$

These conditions justify the Riemann limit and exchanging limit and expectation. As the mesh of Π_ε tends to zero,

$$\begin{aligned} \overline{\mathcal{L}}_{\text{VDM}}^{[\varepsilon, T]} &:= \mathbb{E}_{\mathbf{x}_0 \sim p_{\text{data}}} \lim_{|\Pi_\varepsilon| \rightarrow 0} \mathcal{L}_{\text{VDM}}^{\Pi_\varepsilon}(\mathbf{x}_0) \\ &= -\frac{1}{2} \int_\varepsilon^T \text{SNR}'(s) \mathbb{E}_{\substack{\mathbf{x}_0 \sim p_{\text{data}} \\ \mathbf{x}_s \sim q(\mathbf{x}_s | \mathbf{x}_0)}} \|\mathbf{x}_0 - \mathbf{x}_\phi(\mathbf{x}_s, s)\|^2 ds. \end{aligned}$$

Sample $\mathbf{x}_s$ directly as $\mathbf{x}_s = \alpha_s \mathbf{x}_0 + \sigma_s \epsilon$. With $s \sim \mathcal{U}[\varepsilon, T]$ and $\epsilon \sim \mathcal{N}(0, I_D)$,

$$\overline{\mathcal{L}}_{\text{VDM}}^{[\varepsilon, T]} = -\frac{T - \varepsilon}{2} \mathbb{E}_{\mathbf{x}_0, s, \epsilon} \left[\text{SNR}'(s) \|\mathbf{x}_0 - \mathbf{x}_\phi(\alpha_s \mathbf{x}_0 + \sigma_s \epsilon, s)\|^2 \right].$$

Practical consequence

Training needs one clean sample, one random time, and one Gaussian noise draw—not a simulated forward trajectory.

Supplement: continuous VDM is weighted DSM

For $s < T$ with $\alpha_s > 0$, define the score induced by $\mathbf{x}_\phi$:

$$s_\phi(\mathbf{x}_s, s) := \frac{\alpha_s \mathbf{x}_\phi(\mathbf{x}_s, s) - \mathbf{x}_s}{\sigma_s^2}.$$

Since

$$\nabla_{\mathbf{x}_s} \log q(\mathbf{x}_s | \mathbf{x}_0) = \frac{\alpha_s \mathbf{x}_0 - \mathbf{x}_s}{\sigma_s^2},$$

the continuous objective is

$$\bar{\mathcal{L}}_{\text{VDM}}^{[\varepsilon, T]} = \frac{1}{2} \int_{\varepsilon}^T \underbrace{-\text{SNR}'(s) \frac{\sigma_s^4}{\alpha_s^2}}_{\text{weight derived from the ELBO}} \mathbb{E} \left\| \nabla_{\mathbf{x}_s} \log q(\mathbf{x}_s | \mathbf{x}_0) - s_\phi(\mathbf{x}_s, s) \right\|^2 ds.$$

Same regression target, principled weight

VDM and DSM learn the same conditional denoising regression. The difference is that VDM derives the time weight from reverse-kernel KL divergences.

Supplement: what VDM adds to Score SDE

Score-SDE training specifies a score field along a probability path. VDM adds a particular probabilistic factorization of that path.

What is additional in VDM?

- **One ELBO for all components:** reconstruction, terminal prior matching, and every reverse transition are trained within a single latent-variable objective.
- **KL-derived weighting:** the relative importance of noise levels follows from the reverse conditionals, rather than from a separately selected DSM weight $\omega(t)$.
- **Encoder-decoder semantics:** the explicit hierarchy supports variational likelihood comparison and bits-back or lossless-compression interpretations.

Connection to sampling

The learned reverse kernels discretize the same reverse dynamics represented continuously by the reverse SDE or PF-ODE.