STOCHASTIC INTERPOLANTS VI

Connections, Algorithms, Numerical Experiments, and Semester Summary

Ziseok Lee

April 30, 2026

This final document connects stochastic interpolants to:

score-based diffusion, stochastic localization, denoising, rectified flows.

It then turns the theory into practical algorithms:

learning b, s, η_z and sampling by ODE/SDE.

It ends with numerical experiments and the semester theorem chain.

Section	content
5.1	score-based diffusion and stochastic localization
5.2	denoising methods and SURE
5.3	rectified flows
6.1	learning objectives and empirical risks
6.2	sampling algorithms
7.1	2D experiments
7.2	128D Gaussian mixture experiments
7.3	image generation and mirror interpolation
8	conclusion and semester synthesis

Score-based diffusion often begins with the Ornstein–Uhlenbeck process:

$$\mathrm{d}Z_\tau = -Z_\tau \,\mathrm{d}\tau + \sqrt{2}\,\mathrm{d}W_\tau, \qquad Z_0 = x_1 \sim \rho_1.$$

The solution is:

$$Z_\tau = e^{-\tau} x_1 + \sqrt{2} \int_0^\tau e^{-\tau+s}\,\mathrm{d}W_s.$$

The stochastic integral is Gaussian with covariance:

$$2\int_0^\tau e^{-2(\tau-s)}\,\mathrm{d}s = 1 - e^{-2\tau}.$$

Thus:

$$Z_\tau \sim e^{-\tau} x_1 + \sqrt{1 - e^{-2\tau}}\, z.$$

Set:

$$t = e^{-\tau}.$$

Then:

$$\tau = -\log t,$$

and:

$$Z_{-\log t} = tx_1 + \sqrt{1-t^2}\, z.$$

This is a one-sided spatially linear interpolant:

$$\boxed{x_t^{os,lin} = \alpha(t)z + \beta(t)x_1}$$

with:

$$\alpha(t) = \sqrt{1-t^2}, \qquad \beta(t) = t.$$

The one-sided interpolant:

$$x_t = \sqrt{1 - t^2}\, z + t x_1$$

satisfies:

$$x_0 = z \sim \mathcal{N}(0, I_d),$$
$$x_1 = x_1 \sim \rho_1.$$

Thus finite-time stochastic interpolants recover the OU noising marginals without needing:

$$\tau \in [0, \infty).$$

The infinite-time diffusion variable has become:

$$t \in [0, 1].$$

For:

$$\alpha(t) = \sqrt{1-t^2}, \qquad \beta(t) = t,$$

we have:

$$\dot{\alpha}(t) = -\frac{t}{\sqrt{1-t^2}}, \qquad \dot{\beta}(t) = 1.$$

Thus:

$$b(t,x) = -\frac{t}{\sqrt{1-t^2}}\eta_z^{os}(t,x) + \eta_1^{os}(t,x).$$

Since:

$$s(t,x) = -\alpha^{-1}\eta_z^{os} = -\frac{1}{\sqrt{1-t^2}}\eta_z^{os},$$

we get:

$$\boxed{b(t,x) = t\, s(t,x) + \eta_1^{os}(t,x).}$$

Even though:

$$\dot{\alpha}(t) = -\frac{t}{\sqrt{1-t^2}}$$

is singular as $t \to 1$, the velocity:

$$b(t,x) = t\, s(t,x) + \eta_1^{os}(t,x)$$

is well-behaved under the interpolant formulation.

The algebraic constraint:

$$x = \sqrt{1-t^2}\eta_z^{os} + t\eta_1^{os}$$

cancels the apparent singularity.

This is a key advantage of constructing the marginal path first.

If instead one time-changes the OU SDE directly, define:

$$Z_t^B = Z_{\tau=-\log t}.$$

Then the backward SDE contains:

$$\mathrm{d}Z_t^B = t^{-1} Z_t^B \,\mathrm{d}t + \sqrt{2t^{-1}}\,\mathrm{d}W_t^B.$$

The coefficients blow up at:

$$t = 0.$$

The corresponding forward SDE also inherits singular coefficients.

This makes direct finite-time compression of OU dynamics analytically and numerically difficult.

Stochastic interpolants separate:

density path from sampler.

The OU-inspired path:

$$x_t = \sqrt{1 - t^2} z + t x_1$$

defines $\rho(t)$.

Then one may choose any:

$$\varepsilon(t) \geq 0$$

and sample using:

$$\mathrm{d}X_t = (b + \varepsilon s)(t, X_t)\,\mathrm{d}t + \sqrt{2\varepsilon(t)}\,\mathrm{d}W_t.$$

The sampler need not be the time-changed OU process.

Stochastic localization also studies evolving probability laws by progressively revealing data through noise.

The SBDM one-sided interpolant:

$$x_t = \sqrt{1 - t^2} z + t x_1$$

is a finite-time representation of the same Gaussian revelation principle.

The stochastic interpolant perspective emphasizes:

$$\text{the path of marginals } \rho(t),$$

not necessarily the original forward noising process.

For a one-sided spatially linear interpolant:

$$x_t = \alpha(t)z + \beta(t)x_1.$$

Assume:

$$\beta(t) \neq 0.$$

Then:

$$x_1 = \beta^{-1}(t)(x_t - \alpha(t)z).$$

Conditioning on x_t:

$$\boxed{\mathbb{E}[x_1 \mid x_t] = \beta^{-1}(t)\left(x_t - \alpha(t)\eta_z^{os}(t, x_t)\right).}$$

This is the Bayes-optimal denoiser.

The identity:

$$\mathbb{E}[x_1 \mid x_t] = \beta^{-1}(x_t - \alpha\eta_z)$$

is a Stein-type unbiased risk estimator form.

It shows that the denoiser:

$$\eta_z = \mathbb{E}[z \mid x_t]$$

contains the same information as the clean-data predictor:

$$\mathbb{E}[x_1 \mid x_t].$$

Thus:

$$\text{noise prediction} \iff \text{clean-data prediction} \iff \text{score prediction}.$$

LEMMA 44: CONDITIONAL PREDICTION OF x_s

For:

$$x_t = \alpha(t)z + \beta(t)x_1,$$

and $s, t \in [0, 1], t > 0$, we have:

$$\boxed{\mathbb{E}[x_s \mid x_t] = \frac{\beta(s)}{\beta(t)}x_t + \left(\alpha(s) - \frac{\alpha(t)\beta(s)}{\beta(t)}\right)\eta_z^{os}(t, x_t).}$$

At $t = 0$:

$$\mathbb{E}[x_s \mid x_0] = \alpha(s)x_0 + \beta(s)\mathbb{E}[x_1].$$

Start with:

$$x_s = \alpha(s)z + \beta(s)x_1.$$

Use:

$$x_1 = \beta^{-1}(t)(x_t - \alpha(t)z).$$

Then:

$$x_s = \alpha(s)z + \beta(s)\beta^{-1}(t)(x_t - \alpha(t)z).$$

Collect terms:

$$x_s = \frac{\beta(s)}{\beta(t)}x_t + \left(\alpha(s) - \frac{\alpha(t)\beta(s)}{\beta(t)}\right)z.$$

Condition on x_t and use:

$$\mathbb{E}[z \mid x_t] = \eta_z^{os}(t, x_t).$$

Let:

$$t_j = j/N.$$

Start:

$$X_1^{den} = z.$$

Define:

$$X_{j+1}^{den} = \frac{\beta(t_{j+1})}{\beta(t_j)} X_j^{den} + \left(\alpha(t_{j+1}) - \frac{\alpha(t_j)\beta(t_{j+1})}{\beta(t_j)}\right) \eta_z^{os}(t_j, X_j^{den}).$$

This is obtained by replacing:

$$\mathbb{E}[x_{t_{j+1}} \mid x_{t_j}]$$

with a deterministic update.

As:

$$N, j \to \infty, \qquad j/N \to t,$$

the iterated denoising scheme converges to the probability-flow ODE:

$$\boxed{\dot{X}_t = \frac{\dot{\beta}(t)}{\beta(t)} X_t + \left(\dot{\alpha}(t) - \frac{\alpha(t)\dot{\beta}(t)}{\beta(t)} \right) \eta_z^{os}(t, X_t).}$$

This is exactly the one-sided velocity formula from Document V.

If:

$$X_0 = z \sim \mathcal{N}(0, I_d),$$

then:

$$X_1 \sim \rho_1.$$

Let:

$$h = t_{j+1} - t_j.$$

Taylor expand:

$$\frac{\beta(t_{j+1})}{\beta(t_j)} = 1 + \frac{\dot{\beta}(t_j)}{\beta(t_j)} h + O(h^2).$$

Also:

$$\alpha(t_{j+1}) - \alpha(t_j)\frac{\beta(t_{j+1})}{\beta(t_j)} = \left(\dot{\alpha}(t_j) - \alpha(t_j)\frac{\dot{\beta}(t_j)}{\beta(t_j)}\right) h + O(h^2).$$

Substitute into the update:

$$X_{j+1}^{den} = X_j^{den} + h\frac{\dot{\beta}(t_j)}{\beta(t_j)}X_j^{den}$$

$$+h\left(\dot{\alpha}(t_j) - \alpha(t_j)\frac{\dot{\beta}(t_j)}{\beta(t_j)}\right)\eta_z^{os}(t_j, X_j^{den}) + O(h^2).$$

Thus:

$$\frac{X_{j+1}^{den} - X_j^{den}}{h} = \frac{\dot{\beta}}{\beta}X_j^{den} + \left(\dot{\alpha} - \frac{\alpha\dot{\beta}}{\beta}\right)\eta_z^{os} + O(h).$$

Let $h \to 0$ to get the ODE.

Suppose the original probability-flow ODE is learned exactly:

$$\frac{\mathrm{d}}{\mathrm{d}t} X_t(x) = b(t, X_t(x)), \qquad X_0(x) = x.$$

Let:

$$X_1(x)$$

be the terminal transport map.

Define a new interpolant:

$$\boxed{x_t^{rec} = \alpha(t)x_0 + \beta(t)X_1(x_0).}$$

Endpoint check:

$$x_0^{rec} = x_0, \qquad x_1^{rec} = X_1(x_0) \sim \rho_1.$$

Differentiate:

$$x_t^{rec} = \alpha(t)x_0 + \beta(t)X_1(x_0).$$

Then:

$$\dot{x}_t^{rec} = \dot{\alpha}(t)x_0 + \dot{\beta}(t)X_1(x_0).$$

The rectified velocity is:

$$\boxed{b^{rec}(t,x) = \mathbb{E}[\dot{\alpha}(t)x_0 + \dot{\beta}(t)X_1(x_0) \mid x_t^{rec} = x].}$$

It minimizes:

$$L_b^{rec}[\hat{b}] = \int \mathbb{E}\left[\frac{1}{2}|\hat{b}(t,x_t^{rec})|^2 - (\dot{\alpha}x_0 + \dot{\beta}X_1(x_0)) \cdot \hat{b}(t,x_t^{rec})\right] \mathrm{d}t.$$

Define:

$$M(t, x) = \alpha(t)x + \beta(t)X_1(x).$$

Assume for every t:

$$M(t, \cdot)$$

is invertible.

Let:

$$N(t, \cdot) = M(t, \cdot)^{-1}.$$

Then:

$$N(t, M(t, x)) = x.$$

This allows a closed form for the rectified velocity.

Under the invertibility assumption:

$$\boxed{b^{rec}(t,x) = \dot{\alpha}(t)N(t,x) + \dot{\beta}(t)X_1(N(t,x)).}$$

The probability-flow ODE:

$$\dot{X}_t^{rec} = b^{rec}(t, X_t^{rec}), \qquad X_0^{rec} = x$$

has solution:

$$\boxed{X_t^{rec}(x) = \alpha(t)x + \beta(t)X_1(x).}$$

At $t = 1$:

$$X_1^{rec}(x) = X_1(x).$$

Write:

$$x_t^{rec} = M(t, z).$$

Then:

$$N(t, x_t^{rec}) = N(t, M(t, z)) = z.$$

Therefore:

$$\begin{aligned} b^{rec}(t, x_t^{rec}) &= \dot{\alpha}(t) N(t, M(t, z)) + \dot{\beta}(t) X_1(N(t, M(t, z))) \\ &= \dot{\alpha}(t) z + \dot{\beta}(t) X_1(z) = \dot{x}_t^{rec}. \end{aligned}$$

Thus the stated b^{rec} is the conditional velocity.

Set:

$$X_t^{rec}(x) = M(t,x) = \alpha(t)x + \beta(t)X_1(x).$$

Then:

$$b^{rec}(t, X_t^{rec}(x)) = b^{rec}(t, M(t,x)).$$

Using the velocity formula:

$$b^{rec}(t, M(t,x)) = \dot{\alpha}(t)x + \dot{\beta}(t)X_1(x).$$

But:

$$\frac{\mathrm{d}}{\mathrm{d}t}M(t,x) = \dot{\alpha}(t)x + \dot{\beta}(t)X_1(x).$$

So X_t^{rec} solves the rectified ODE.

The terminal map is:

$$X_1^{rec} = X_1.$$

Thus rectification changes the intermediate path:

$$t \mapsto X_t^{rec},$$

but not the final generative map:

$$x_0 \mapsto x_1.$$

For:

$$\alpha(t) = 1 - t, \qquad \beta(t) = t,$$

the rectified paths are straight:

$$X_t^{rec}(x) = (1 - t)x + tX_1(x).$$

Straight paths alone do not imply optimal transport.

Optimal transport requires more than straight trajectories.

For quadratic cost, the optimal map has gradient structure:

$$T = \nabla \phi$$

for a convex potential ϕ.

Rectification without enforcing:

$$b^{rec} = \nabla \phi_t$$

does not necessarily produce the optimal map.

Thus:

$$\text{straight paths} \quad \not\Rightarrow \quad \text{optimal transport.}$$

Training draws:

$$t_i \sim \mathrm{Unif}[0,1],$$
$$(x_0^i, x_1^i) \sim \nu,$$
$$z_i \sim \mathcal{N}(0, I_d).$$

Construct:

$$x_{t_i}^i = I(t_i, x_0^i, x_1^i) + \gamma(t_i) z_i.$$

All objectives are Monte Carlo estimates of population quadratic losses.

The deterministic velocity loss is:

$$\widehat{L}_b(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{2}|b_\theta(t_i, x_{t_i}^i)|^2 - b_\theta(t_i, x_{t_i}^i)\cdot(\partial_t I_i + \dot{\gamma}_i z_i)\right].$$

With exact population expectation:

$$b_\theta = b$$

is the unique minimizer.

This requires no simulation of the ODE during training.

Instead of b, learn:

$$v(t,x) = \mathbb{E}[\partial_t I \mid x_t = x].$$

Empirical objective:

$$\widehat{L}_v(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{2}|v_\theta(t_i, x^i_{t_i})|^2 - v_\theta(t_i, x^i_{t_i})\cdot \partial_t I(t_i, x^i_0, x^i_1)\right].$$

Then:

$$\hat{b} = \hat{v} - \gamma\dot{\gamma}\hat{s}.$$

The denoiser loss is:

$$\widehat{L}_{\eta_z}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{2}|\eta_\theta(t_i, x_{t_i}^i)|^2 - \eta_\theta(t_i, x_{t_i}^i)\cdot z_i\right].$$

Then:

$$\hat{s}(t,x) = -\gamma^{-1}(t)\eta_\theta(t,x).$$

This is preferred when the score objective:

$$\gamma^{-1}z$$

is numerically unstable.

The score loss is:

$$\widehat{L}_s(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{2}|s_\theta(t_i, x_{t_i}^i)|^2 + \gamma^{-1}(t_i) z_i \cdot s_\theta(t_i, x_{t_i}^i)\right].$$

Its population minimizer is:

$$s = \nabla \log \rho.$$

In practice, avoid sampling t_i too close to endpoints unless the singularity is controlled.

Because the score loss may have endpoint variance, use:

$$z \quad \text{and} \quad -z.$$

Construct:

$$x_t^+ = I(t, x_0, x_1) + \gamma(t)z,$$
$$x_t^- = I(t, x_0, x_1) - \gamma(t)z.$$

Antithetic averaging cancels odd fluctuations and improves stability.

This is especially useful when:

$$\gamma(t) \to 0$$

near endpoints.

Let $F(x) = z \cdot s(t, x)$. Then:

$$F(I + \gamma z) - F(I - \gamma z) = 2\gamma z^\top \nabla s(t, I) z + O(\gamma^3).$$

Therefore:

$$\frac{z \cdot s(t, x_t^+) - z \cdot s(t, x_t^-)}{2\gamma} = z^\top \nabla s(t, I) z + O(\gamma^2).$$

This removes the leading singular behavior associated with division by γ.

Given learned:

$$\hat{b}(t, x),$$

sample:

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{X}_t = \hat{b}(t, \hat{X}_t), \qquad \hat{X}_0 \sim \rho_0.$$

Numerical integrators:

Euler, Runge–Kutta, adaptive ODE solvers.

ODE sampling gives deterministic maps:

$$\hat{X}_1 = \hat{\Phi}_{0,1}(\hat{X}_0).$$

It also supports exact likelihood estimation via divergence integration.

Given:

$$\hat{b}, \qquad \hat{s},$$

define:

$$\hat{b}^F = \hat{b} + \varepsilon \hat{s}.$$

Sample:

$$\mathrm{d}\hat{X}_t^F = \hat{b}^F(t, \hat{X}_t^F)\,\mathrm{d}t + \sqrt{2\varepsilon(t)}\,\mathrm{d}W_t.$$

Euler–Maruyama step:

$$\hat{X}_{t+h} = \hat{X}_t + h\hat{b}^F(t, \hat{X}_t) + \sqrt{2\varepsilon(t)h}\,\xi_t,$$

where:

$$\xi_t \sim \mathcal{N}(0, I_d).$$

The likelihood bound from Document III is:

$$\mathrm{KL}(\rho_1 \| \hat{\rho}(1)) \leq \frac{1}{2\varepsilon} \Delta L_b + \frac{\varepsilon}{2} \Delta L_s.$$

The optimal constant in this upper bound is:

$$\varepsilon^* = \left(\frac{\Delta L_b}{\Delta L_s} \right)^{1/2}.$$

Interpretation:

$$\Delta L_b \ll \Delta L_s \Rightarrow \varepsilon^* \text{ small},$$
$$\Delta L_s \ll \Delta L_b \Rightarrow \varepsilon^* \text{ large}.$$

For:

$$x_t = \alpha(t)z + \beta(t)x_1,$$

learn:

$$\hat{\eta}_z.$$

Construct:

$$\hat{s} = -\alpha^{-1}\hat{\eta}_z,$$
$$\hat{b} = \frac{\dot{\beta}}{\beta}x + \left(\dot{\alpha} - \frac{\alpha\dot{\beta}}{\beta}\right)\hat{\eta}_z.$$

Then:

$$\hat{b}^F = \hat{b} + \varepsilon\hat{s}.$$

Integrate ODE or SDE from:

$$X_0 \sim \mathcal{N}(0, I_d)$$

to $t = 1$.

When sampling with a denoiser:

$$\hat{s} = -\hat{\eta}_z/\gamma,$$

singularities may arise near:

$$t = 1.$$

A practical procedure:

integrate only to $t_f < 1$,

then use the denoising formula:

$$\mathbb{E}[x_1 \mid x_{t_f}] = \beta^{-1}(t_f)\left(x_{t_f} - \alpha(t_f)\hat{\eta}_z(t_f, x_{t_f})\right).$$

This is the finite-step version of the SURE identity.

If:

$$\partial_t \hat{\rho} + \nabla \cdot (\hat{b}\hat{\rho}) = 0,$$

then:

$$\log \hat{\rho}(1, x_1) = \log \rho_0(X_{1,0}(x_1)) - \int_0^1 \nabla \cdot \hat{b}(t, X_{1,t}(x_1))\, \mathrm{d}t.$$

Thus:

$$H(\rho_1 \| \hat{\rho}(1)) = \mathbb{E}_{x_1 \sim \rho_1} \int_0^1 \nabla \cdot \hat{b}(t, X_{1,t}(x_1))\, \mathrm{d}t - \mathbb{E}_1 \log \rho_0(X_{1,0}(x_1)).$$

High-dimensional divergence:

$$\nabla \cdot b(t,x) = \mathrm{Tr}(\nabla_x b(t,x)).$$

Hutchinson estimator:

$$\mathrm{Tr}(A) = \mathbb{E}_\xi[\xi^\top A \xi],$$

where:

$$\mathbb{E}[\xi\xi^\top] = I_d.$$

Thus:

$$\nabla \cdot b(t,x) = \mathbb{E}_\xi[\xi^\top \nabla_x b(t,x)\xi].$$

This makes likelihood estimation feasible for neural velocity fields.

The 2D experiments compare:

$$\gamma(t) = 0, \qquad \gamma(t) = \sqrt{t(1-t)}, \qquad \gamma(t) = t(1-t),$$

and other smooth choices.

They also vary:

$$\varepsilon \in \{0, 0.5, 1.0, 2.5\}.$$

Key qualitative observation:

$$\varepsilon > 0$$

often improves the learned model when b, s are approximate.

But large ε requires smaller numerical steps.

When:

$$\varepsilon = 0,$$

sampling uses the probability-flow ODE.

When:

$$\varepsilon > 0,$$

sampling uses the learned SDE:

$$\mathrm{d}X_t = (\hat{b} + \varepsilon \hat{s})\,\mathrm{d}t + \sqrt{2\varepsilon}\,\mathrm{d}W_t.$$

The density path $\rho(t)$ is fixed by γ, not ε.

Experimentally:

$$\gamma(t) = \sqrt{t(1-t)}$$

gave strong ODE performance because endpoint score information enters b.

The paper compares:

$$\log \rho_1(x)$$

with:

$$\log \hat{\rho}(1, x).$$

Quantities:

$$\mathbb{E}|\log \rho_1 - \log \hat{\rho}(1)|,$$
$$\mathrm{Var}(|\log \rho_1 - \log \hat{\rho}(1)|).$$

The experiments show:

SDE sampling can outperform ODE sampling when fields are imperfect.

Target:

$$\rho_1 = \text{Gaussian mixture in } d = 128$$

with five modes.

Base:

$$\rho_0 = \mathcal{N}(0, I_d).$$

Interpolant:

$$\alpha(t) = 1 - t, \qquad \beta(t) = t, \qquad \gamma(t) = \sqrt{t(1-t)}.$$

Four learning combinations:

$$(v, s), \qquad (v, \eta), \qquad (b, s), \qquad (b, \eta).$$

The model and target are projected to two coordinates.

KDEs estimate:

$$\rho_1(x), \qquad \hat{\rho}_1(x).$$

Monte Carlo KL with control variate:

$$\mathrm{KL}(\rho_1 \| \hat{\rho}_1) \approx \frac{1}{N_e} \sum_{i=1}^{N_e} \left[\log \rho_1(x_i) - \log \hat{\rho}_1(x_i) - \left(\frac{\hat{\rho}_1(x_i)}{\rho_1(x_i)} - 1 \right) \right].$$

The control variate reduces variance and keeps the estimate nonnegative by concavity of $\log$.

Empirical behavior as ε varies:

$$\varepsilon \text{ small} \Rightarrow \text{model overly concentrated in modes},$$

$$\varepsilon \text{ large} \Rightarrow \text{model overly spread into tails}.$$

An intermediate value of ε performs best.

The reported pattern:

$$\text{learning } b \text{ often outperforms learning } v,$$

$$\text{learning } \eta_z \text{ often outperforms learning } s,$$

with endpoint singularities handled carefully.

The paper demonstrates scalability on image generation.

Two examples:

one-sided Gaussian-base interpolants,

mirror interpolants.

One-sided examples:

$$x_t = (1-t)z + tx_1,$$

and:

$$x_t = \cos(\pi t/2)z + \sin(\pi t/2)x_1.$$

Networks:

U-Net parameterizations of $\hat{b}, \hat{s}, \hat{\eta}_z$.

ODE sampling:

$$\varepsilon = 0.$$

SDE sampling:

$$\varepsilon > 0.$$

Using the same initial Gaussian sample, different SDE paths can produce different terminal images.

Increasing ε increases sample diversity from the same initial condition.

This illustrates:

ODE = deterministic map, SDE = stochastic conditional ensemble.

The image experiment includes nearest-neighbor comparison.

Given a generated image:

$$\hat{x},$$

find nearest training images under:

$$\ell^1 \text{ distance}.$$

Visual distinction from nearest neighbors supports that the model is not merely memorizing.

Mathematically, this is not a theorem, but a useful empirical diagnostic.

For the mirror interpolant:

$$x_t = x_1 + \gamma(t) z.$$

The velocity is:

$$b = \dot{\gamma}\, \eta_z.$$

ODE sampling tends to preserve the input.

SDE sampling:

$$\mathrm{d}X_t = (\dot{\gamma}\eta_z + \varepsilon s)\,\mathrm{d}t + \sqrt{2\varepsilon}\,\mathrm{d}W_t$$

can resample to a nearby but different image.

This demonstrates data-distribution-preserving stochastic transformations.

Training:

$$(t_i, x_0^i, x_1^i, z_i) \mapsto x_{t_i}^i \mapsto \widehat{L}_b, \widehat{L}_s, \widehat{L}_{\eta_z}.$$

Sampling:

$$\hat{b}, \hat{s} \mapsto \hat{b}^F = \hat{b} + \varepsilon \hat{s} \mapsto \text{ODE/SDE integration.}$$

Validation:

cross-entropy, KL on tractable projections, sample quality diagnostics.

$$x_t = I(t, x_0, x_1) + \gamma(t)z$$

$$\Downarrow$$

$$\rho(t) = \mathcal{L}(x_t), \qquad \rho(0) = \rho_0, \qquad \rho(1) = \rho_1$$

$$\Downarrow$$

$$\partial_t \rho + \nabla \cdot (b\rho) = 0$$

$$\Downarrow$$

$$b = \mathbb{E}[\dot{x}_t \mid x_t = x], \qquad s = -\gamma^{-1}\mathbb{E}[z \mid x_t = x]$$

$$\Downarrow$$

b, s, η_z are quadratic-regression minimizers.

$$b^F = b + \varepsilon s, \qquad b^B = b - \varepsilon s$$

$$\Downarrow$$

$$\partial_t \rho + \nabla \cdot (b^F \rho) = \varepsilon \Delta \rho$$

$$\Downarrow$$

$$\mathrm{d}X_t^F = b^F(t, X_t^F)\,\mathrm{d}t + \sqrt{2\varepsilon}\,\mathrm{d}W_t$$

$$\Downarrow$$

$$X_0^F \sim \rho_0 \quad \Rightarrow \quad X_1^F \sim \rho_1.$$

With learned fields:

$$\mathrm{KL}(\rho_1 \| \hat{\rho}(1)) \leq \frac{1}{2\varepsilon} \Delta L_b + \frac{\varepsilon}{2} \Delta L_s.$$

1. Generative modeling is probability transport.
2. The interpolant specifies a density path.
3. ODEs and SDEs can sample the same path.
4. Velocity and score are conditional expectations.
5. Quadratic losses are not heuristics; they are projection formulas.
6. Diffusion gives KL control through Fisher dissipation.

By the end of Semester I, students should be able to prove:

smoothness and positivity of $\rho(t)$,

$$\partial_t \rho + \nabla \cdot (b\rho) = 0,$$

$$s = -\gamma^{-1}\mathbb{E}[z \mid x_t = x],$$

b, s, η_z minimize quadratic risks,

$$\partial_t \rho + \nabla \cdot ((b + \varepsilon s)\rho) = \varepsilon \Delta \rho,$$

KL control for learned SDEs,

density and cross-entropy estimation formulas.

Possible proof-based final projects:

1. Reprove Theorem 6 with alternative regularity assumptions.

2. Analyze endpoint singularities for different $\gamma(t)$.

3. Implement and compare $(b, s), (b, \eta), (v, s), (v, \eta)$.

4. Derive the Gaussian mixture formulas in full detail.

5. Study the Schrödinger bridge max-min problem numerically.

6. Extend one-sided interpolants to non-isotropic Gaussian bases.

Semester I focused on:

$\mathbb{R}^d$, densities, ODEs/SDEs, transport and Fokker–Planck equations.

Semester II abstracts this to:

general state spaces, Markov generators, jump processes, discrete diffusion, manifolds and multimodal spaces.

The object replacing b and s is:

$$\boxed{\mathcal{L}_t}$$

the infinitesimal generator.

Stochastic interpolants provide:

a finite-time, proof-level unification of flows and diffusions.

The framework begins with:

$$x_t = I(t, x_0, x_1) + \gamma(t)z,$$

and ends with:

trainable ODE/SDE generative models with likelihood theory.

The semester's central insight:

Do not start from an algorithm. Start from a probability path and prove its dynamics.

This prepares the general-state-space viewpoint of Semester II.