STOCHASTIC INTERPOLANTS I

Mathematical Preliminaries, Background, Motivation, and Related Work

Ziseok Lee

April 30, 2026

Semester I is a proof-level reading of the stochastic interpolants framework.

This first document prepares the mathematical language needed for the rest of the semester:

$$\text{probability} + \text{analysis} + \text{PDE} + \text{stochastic calculus} \quad \Longrightarrow \quad \text{generative modeling.}$$

The guiding thesis is:

> A generative model is a law-preserving dynamical realization of a probability path.

The stochastic interpolant framework separates:

> designing the path of densities $\rho(t)$ from choosing an ODE or SDE that samples it.

Given two probability densities

$$\rho_0, \rho_1 : \mathbb{R}^d \to [0, \infty), \qquad \int_{\mathbb{R}^d} \rho_i(x)\,\mathrm{d}x = 1,$$

construct a process that transports

$$x_0 \sim \rho_0 \qquad \rightsquigarrow \qquad x_1 \sim \rho_1.$$

The course goal is not merely to implement such a process. The goal is to prove:

existence of the interpolating density $\rho(t, x)$,

PDEs satisfied by $\rho(t, x)$,

ODE/SDE samplers with the same marginals,

training objectives and their minimizers,

likelihood, cross-entropy, and KL control.

1. Mathematical preliminaries, background, motivation, and related work.
2. Stochastic interpolant definitions, assumptions, transport equations, score, quadratic objectives.
3. Generative models, likelihood control, density estimation, cross-entropy.
4. Instantiations and extensions: diffusive, one-sided, mirror, Schrödinger bridge.
5. Spatially linear interpolants: factorization, design choices, latent variables, diffusion coefficients.
6. Connections, algorithms, numerical experiments, and semester synthesis.

The proof chain throughout the semester is:

$$x_t \implies \rho(t) \implies \partial_t \rho + \nabla \cdot (b\rho) = 0 \implies \text{ODE/SDE samplers} \implies \text{trainable objectives}.$$

MAIN PAPER ORGANIZATION

The semester follows the paper organization closely:

Section 1	:	motivation, contributions, related work, notation,
Section 2	:	general stochastic interpolant framework,
Section 3	:	instantiations and extensions,
Section 4	:	spatially linear interpolants,
Section 5	:	connections with other methods,
Section 6	:	algorithmic aspects,
Section 7	:	numerical experiments,
Appendix B	:	proofs.

Pedagogical rule:

> No algorithm without its density equation and proof.

A generative model aims to sample from a complex unknown distribution:

$$x \sim \rho_{\text{data}}.$$

We usually have samples but not a tractable density.

The modern continuous-time viewpoint is:

$$z \sim \rho_0 \text{ easy} \quad \rightsquigarrow \quad x \sim \rho_1 \text{ data}$$

by solving a dynamical system.

Examples:

$$\begin{aligned}
\text{normalizing flow} &: \quad x_1 = T(x_0), \\
\text{continuous normalizing flow} &: \quad \dot{X}_t = b(t, X_t), \\
\text{diffusion model} &: \quad \mathrm{d}X_t = b(t, X_t)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}W_t, \\
\text{stochastic interpolant} &: \quad x_t = I(t, x_0, x_1) + \gamma(t)z.
\end{aligned}$$

MOTIVATION: WHY A RIGOROUS COURSE?

Many courses present generative modeling as:

$$\text{architecture} + \text{loss} + \text{sampler} + \text{FID}.$$

This course instead asks:

What mathematical theorem justifies the loss and sampler?

For every method we want proofs of:

the probability path,

the PDE governing the path,

the ODE/SDE generating the same marginals,

the regression objective and its unique minimizer,

the statistical consequence of approximation error.

The framework is powerful because the same object can be read as:

$$\text{transport map} = \text{PDE} = \text{stochastic process} = \text{optimization problem}.$$

A classical score-based diffusion often runs through an infinite-time noising limit:

$$x_\tau = e^{-\tau} x_1 + \sqrt{1 - e^{-2\tau}}\, z, \qquad \tau \in [0, \infty).$$

The stochastic interpolant idea instead works on:

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

Prototype:

$$x_t = (1 - t)x_0 + t x_1 + \sqrt{2t(1 - t)}\, z.$$

Endpoint check:

$$x_0 = x_0, \qquad x_1 = x_1,$$

because the latent Gaussian term vanishes at both endpoints.

Core question:

What is the law of x_t, and how can we sample it without knowing x_1?

The stochastic interpolant defines a path of laws:

$$\rho(t) = \mathcal{L}(x_t).$$

Once $\rho(t)$ is known through its evolution equation, we can sample the same path in different ways:

$$\dot{X}_t = b(t, X_t),$$

or

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

These have different sample paths but the same marginal densities:

$$\mathcal{L}(X_t) = \rho(t).$$

This is the main conceptual separation:

$$\boxed{\text{interpolant path } \rho(t) \neq \text{particular sampler path } X_t.}$$

For a random variable X with density ρ,

$$X \sim \rho \quad \Longleftrightarrow \quad \mathbb{P}(X \in A) = \int_A \rho(x)\,\mathrm{d}x.$$

The law of X is denoted

$$\mathcal{L}(X).$$

A time-dependent density satisfies

$$\rho : [0,1] \times \mathbb{R}^d \to [0,\infty), \qquad \int_{\mathbb{R}^d} \rho(t,x)\,\mathrm{d}x = 1.$$

For a test function ϕ,

$$\mathbb{E}[\phi(X_t)] = \int_{\mathbb{R}^d} \phi(x)\rho(t,x)\,\mathrm{d}x.$$

The paper uses standard spaces:

$$C^k(\mathbb{R}^d) = \{k\text{-times continuously differentiable functions}\}.$$

$$C_0^k(\mathbb{R}^d) = \{C^k \text{ compactly supported functions}\}.$$

For vector-valued functions:

$$(C^k(\mathbb{R}^d))^d.$$

For time-dependent functions:

$$b \in C^0([0,1]; (C^p(\mathbb{R}^d))^d)$$

means $t \mapsto b(t, \cdot)$ is continuous into the C^p-topology.

Why this matters:

PDE identities first hold weakly, then regularity promotes them.

A test function is typically

$$\phi \in C_0^\infty(\mathbb{R}^d).$$

It probes distributions through:

$$\mu \mapsto \int_{\mathbb{R}^d} \phi(x)\mu(\,\mathrm{d}x).$$

Two probability measures μ, ν agree if

$$\int \phi \,\mathrm{d}\mu = \int \phi \,\mathrm{d}\nu \qquad \forall \phi \in C_0^\infty(\mathbb{R}^d).$$

Thus, to prove a PDE weakly, it suffices to prove:

$$\frac{\mathrm{d}}{\mathrm{d}t} \int \phi\rho = \text{right-hand side tested against } \phi.$$

The identity

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

is equivalent in weak form to

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathbb{R}^d} \phi(x)\rho(t,x)\,\mathrm{d}x = \int_{\mathbb{R}^d} \nabla\phi(x) \cdot b(t,x)\rho(t,x)\,\mathrm{d}x.$$

Proof by integration by parts:

$$\int \phi\, \partial_t \rho = -\int \phi\, \nabla \cdot (b\rho) = \int \nabla\phi \cdot b\rho.$$

The compact support of ϕ kills boundary terms.

Let X be a random variable with law μ. For an integrable random variable Y,

$$\mathbb{E}[Y \mid X = x]$$

is the function $g(x)$ satisfying:

$$\int \phi(x) g(x) \mu(\,\mathrm{d}x) = \mathbb{E}[\phi(X) Y] \qquad \forall \phi \in C_0^\infty(\mathbb{R}^d).$$

In the stochastic interpolant paper:

$$\mathbb{E}[f(t, x_0, x_1, z) \mid x_t = x]$$

is defined by:

$$\int \phi(x) \mathbb{E}[f \mid x_t = x] \rho(t, x)\,\mathrm{d}x = \mathbb{E}[\phi(x_t) f].$$

Formally, using Dirac deltas,

$$\mathbb{E}[f \mid x_t = x] = \frac{\mathbb{E}[f\,\delta(x - x_t)]}{\mathbb{E}[\delta(x - x_t)]}.$$

Since

$$\rho(t, x) = \mathbb{E}[\delta(x - x_t)],$$

we write:

$$\mathbb{E}[f \mid x_t = x]\rho(t, x) = \mathbb{E}[f\,\delta(x - x_t)].$$

This heuristic is useful for deriving:

$$b(t, x) = \mathbb{E}[\dot{x}_t \mid x_t = x],$$

and

$$s(t, x) = \nabla \log \rho(t, x) = -\gamma(t)^{-1}\mathbb{E}[z \mid x_t = x].$$

A coupling of ρ_0 and ρ_1 is a probability measure

$$\nu(\,\mathrm{d}x_0,\,\mathrm{d}x_1)$$

on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals:

$$\nu(\,\mathrm{d}x_0, \mathbb{R}^d) = \rho_0(x_0)\,\mathrm{d}x_0,$$
$$\nu(\mathbb{R}^d,\,\mathrm{d}x_1) = \rho_1(x_1)\,\mathrm{d}x_1.$$

The independent coupling is:

$$\nu(\,\mathrm{d}x_0,\,\mathrm{d}x_1) = \rho_0(x_0)\rho_1(x_1)\,\mathrm{d}x_0\,\mathrm{d}x_1.$$

A nontrivial coupling can encode geometry, pairing samples from ρ_0 and ρ_1.

The stochastic interpolant includes:

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

Why Gaussian?

$$\mathbb{E}e^{ik\cdot\gamma z} = e^{-\frac{1}{2}\gamma^2|k|^2}.$$

This factor provides spatial smoothing:

$$\rho(t) = \mu_I(t) * \mathcal{N}(0, \gamma^2(t) I_d),$$

where $\mu_I(t) = \mathcal{L}(I(t, x_0, x_1))$.

Thus, for $0 < t < 1$ and $\gamma(t) > 0$,

$$\rho(t, \cdot) \in C^\infty(\mathbb{R}^d).$$

For a random variable $X \in \mathbb{R}^d$, define:

$$g(k) = \mathbb{E}e^{ik\cdot X}.$$

If X has density ρ, then:

$$g(k) = \int_{\mathbb{R}^d} e^{ik\cdot x}\rho(x)\,\mathrm{d}x.$$

Fourier inversion:

$$\rho(x) = (2\pi)^{-d}\int_{\mathbb{R}^d} e^{-ik\cdot x}g(k)\,\mathrm{d}k,$$

provided $g \in L^1$.

For derivatives:

$$\partial_{x_j}\rho = (2\pi)^{-d}\int (-ik_j)e^{-ik\cdot x}g(k)\,\mathrm{d}k.$$

If

$$g(t,k) = g_0(t,k)e^{-\frac{1}{2}\gamma^2(t)|k|^2},$$

then for any $p \in \mathbb{N}$,

$$\int_{\mathbb{R}^d} |k|^p |g(t,k)| \, \mathrm{d}k < \infty \qquad \text{for } 0 < t < 1.$$

Hence:

$$\rho(t,\cdot) \in C^p(\mathbb{R}^d) \qquad \forall p.$$

This is the analytic reason the latent variable $\gamma(t)z$ regularizes the interpolant.

Let

$$\varphi_\gamma(x) = (2\pi\gamma^2)^{-d/2} \exp\left(-\frac{|x|^2}{2\gamma^2}\right).$$

If

$$X = Y + \gamma z, \qquad z \sim \mathcal{N}(0, I_d),$$

then

$$\rho_X = \rho_Y * \varphi_\gamma.$$

Explicitly:

$$\rho_X(x) = \int_{\mathbb{R}^d} \varphi_\gamma(x-y)\rho_Y(y)\,\mathrm{d}y.$$

If ρ_Y is merely a measure,

$$\rho_X(x) = \int \varphi_\gamma(x-y)\mu_Y(\,\mathrm{d}y) > 0.$$

For a strictly positive density $\rho(t, x)$, the score is:

$$s(t, x) = \nabla \log \rho(t, x) = \frac{\nabla \rho(t, x)}{\rho(t, x)}.$$

The score transforms a Laplacian into divergence form:

$$\Delta \rho = \nabla \cdot (\nabla \rho) = \nabla \cdot (s\rho).$$

This identity is the key algebra behind:

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

being equivalent to

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

The Fisher information of ρ is:

$$\int_{\mathbb{R}^d} |\nabla \log \rho(x)|^2 \rho(x)\, \mathrm{d}x.$$

The relative Fisher divergence is:

$$\mathrm{FI}(\rho \| \hat{\rho}) = \int_{\mathbb{R}^d} |\nabla \log \rho - \nabla \log \hat{\rho}|^2 \rho \, \mathrm{d}x.$$

It appears in likelihood control:

$$\mathrm{KL}(\rho(1) \| \hat{\rho}(1)) = \text{drift-error term} - \varepsilon \int_0^1 \mathrm{FI}(\rho(t) \| \hat{\rho}(t))\, \mathrm{d}t.$$

For densities $\rho, \hat{\rho}$,

$$\mathrm{KL}(\rho\|\hat{\rho}) = \int_{\mathbb{R}^d} \rho(x) \log \frac{\rho(x)}{\hat{\rho}(x)} \,\mathrm{d}x.$$

Cross-entropy:

$$\mathrm{H}(\rho\|\hat{\rho}) = -\int_{\mathbb{R}^d} \rho(x) \log \hat{\rho}(x) \,\mathrm{d}x.$$

Entropy:

$$\mathrm{H}(\rho) = -\int \rho \log \rho.$$

Relation:

$$\mathrm{H}(\rho\|\hat{\rho}) = \mathrm{KL}(\rho\|\hat{\rho}) + \mathrm{H}(\rho).$$

The continuity equation is:

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

It expresses conservation of probability mass.

For any region $A \subset \mathbb{R}^d$,

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_A \rho(t,x)\,\mathrm{d}x = -\int_{\partial A} b(t,x)\rho(t,x) \cdot n(x)\,\mathrm{d}S.$$

Velocity b moves mass. It is not necessarily a gradient field.

For $\phi \in C_0^\infty(\mathbb{R}^d)$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[\phi(X_t)] = \frac{\mathrm{d}}{\mathrm{d}t}\int \phi\rho = \int \nabla\phi \cdot b\rho.$$

If X_t solves

$$\dot{X}_t = b(t, X_t),$$

then by the chain rule:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[\phi(X_t)] = \mathbb{E}[\nabla\phi(X_t) \cdot b(t, X_t)].$$

Thus the ODE flow pushes $\rho(0)$ along the continuity equation.

Let $X_{s,t}(x)$ solve:

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

Along a characteristic:

$$\frac{\mathrm{d}}{\mathrm{d}t} \rho(t, X_{s,t}) = \partial_t \rho + b \cdot \nabla \rho.$$

Using

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

gives:

$$\partial_t \rho + b \cdot \nabla \rho = -(\nabla \cdot b)\rho.$$

Therefore:

$$\frac{\mathrm{d}}{\mathrm{d}t} \log \rho(t, X_{s,t}) = -\nabla \cdot b(t, X_{s,t}).$$

Integrating the characteristic identity:

$$\log \rho(t, X_{0,t}(x_0)) = \log \rho_0(x_0) - \int_0^t \nabla \cdot b(\tau, X_{0,\tau}(x_0)) \, \mathrm{d}\tau.$$

Equivalently, for $x_0 = X_{t,0}(x)$:

$$\rho(t, x) = \rho_0(X_{t,0}(x)) \exp\left(- \int_0^t \nabla \cdot b(\tau, X_{t,\tau}(x)) \, \mathrm{d}\tau \right).$$

This is the continuous normalizing flow likelihood formula.

Consider the SDE

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

The density solves:

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

Weak form:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int \phi\rho = \int \nabla\phi \cdot b^F \rho + \varepsilon(t)\int \Delta\phi\,\rho.$$

The right-hand side is generated by Itô's formula.

Let

$$\mathrm{d}X_t = b(t, X_t)\,\mathrm{d}t + \sigma(t, X_t)\,\mathrm{d}W_t.$$

For $f \in C^{1,2}$:

$$\mathrm{d}f(t, X_t) = \partial_t f(t, X_t)\,\mathrm{d}t + \nabla f(t, X_t) \cdot \mathrm{d}X_t + \frac{1}{2}\,\mathrm{Tr}\Big(\sigma\sigma^\top \nabla^2 f(t, X_t)\Big)\,\mathrm{d}t.$$

For

$$\sigma = \sqrt{2\varepsilon}\, I_d,$$

this becomes:

$$\mathrm{d}f = \partial_t f\,\mathrm{d}t + \nabla f \cdot \mathrm{d}X_t + \varepsilon \Delta f\,\mathrm{d}t.$$

For an SDE

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

the generator acts on test functions by:

$$L_t f = b \cdot \nabla f + \varepsilon \Delta f.$$

The adjoint acts on densities:

$$L_t^* \rho = -\nabla \cdot (b\rho) + \varepsilon \Delta \rho.$$

Thus:

$$\partial_t \rho = L_t^* \rho$$

is exactly the Fokker–Planck equation.

A Brownian bridge B_t pinned at 0 at times 0 and 1 satisfies:

$$B_t = W_t - tW_1.$$

Its one-time law is:

$$B_t \sim \mathcal{N}(0, t(1-t)I_d).$$

It also solves:

$$\mathrm{d}B_t = -\frac{B_t}{1-t}\,\mathrm{d}t + \,\mathrm{d}W_t.$$

This later motivates diffusive interpolants:

$$x_t^d = I(t, x_0, x_1) + \sqrt{2a(t)}B_t.$$

Define the backward Brownian motion:

$$W_t^B = -W_{1-t}.$$

A backward SDE is written:

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

It is solved backward from:

$$X_1^B \sim \rho_1.$$

If $Z_\tau = X_{1-\tau}^B$, then Z is a forward-time process.

For a backward SDE

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

the backward Itô formula is:

$$\mathrm{d}f(t, X_t^B) = \partial_t f(t, X_t^B)\,\mathrm{d}t + \nabla f(t, X_t^B)\cdot\,\mathrm{d}X_t^B - \varepsilon(t)\Delta f(t, X_t^B)\,\mathrm{d}t.$$

The Laplacian sign changes because time is reversed.

Score-based diffusion often uses:

$$\mathrm{d}Z_\tau = -Z_\tau \,\mathrm{d}\tau + \sqrt{2}\,\mathrm{d}W_\tau.$$

Solution:

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

As $\tau \to \infty$:

$$Z_\tau \Rightarrow \mathcal{N}(0, I_d).$$

The stochastic interpolant time change:

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

gives:

$$Z_{-\log t} = t x_1 + \sqrt{1 - t^2}\, z.$$

For a density ρ, the Hyvärinen score matching objective is:

$$\int \left(|\hat{s}(x)|^2 + 2\nabla \cdot \hat{s}(x)\right) \rho(x)\, \mathrm{d}x.$$

Using integration by parts:

$$\int \nabla \cdot \hat{s}\, \rho = -\int \hat{s} \cdot \nabla \rho = -\int \hat{s} \cdot s\, \rho.$$

Therefore:

$$\int \left(|\hat{s}|^2 + 2\nabla \cdot \hat{s}\right) \rho = \int |\hat{s} - s|^2 \rho - \int |s|^2 \rho.$$

If

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

then the stochastic interpolant score identity is:

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

Thus learning a denoiser

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

is equivalent to learning the score:

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

This avoids computing $\nabla \cdot \hat{s}$.

Let Y be an L^2 random variable and X another random variable. Then:

$$g^*(x) = \mathbb{E}[Y \mid X = x]$$

is the unique minimizer of:

$$\inf_g \mathbb{E}\left[\frac{1}{2}|g(X)|^2 - Y \cdot g(X)\right].$$

Proof:

$$\mathbb{E}\left[\frac{1}{2}|g(X)|^2 - Y \cdot g(X)\right] = \mathbb{E}\left[\frac{1}{2}|g(X)|^2 - \mathbb{E}[Y \mid X] \cdot g(X)\right].$$

Complete the square:

$$= \frac{1}{2}\mathbb{E}|g(X) - g^*(X)|^2 - \frac{1}{2}\mathbb{E}|g^*(X)|^2.$$

The velocity is a conditional expectation:

$$b(t, x) = \mathbb{E}[\dot{x}_t \mid x_t = x].$$

Therefore it is a least-squares regression target:

$$b = \arg\min_{\hat{b}} \int_0^1 \mathbb{E}\left[\frac{1}{2}|\hat{b}(t, x_t)|^2 - \dot{x}_t \cdot \hat{b}(t, x_t)\right] \mathrm{d}t.$$

The denoiser is also a conditional expectation:

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

Thus:

$$\eta_z = \arg\min_{\hat{\eta}} \int_0^1 \mathbb{E}\left[\frac{1}{2}|\hat{\eta}(t, x_t)|^2 - z \cdot \hat{\eta}(t, x_t)\right] \mathrm{d}t.$$

Normalizing flows learn an invertible map:

$$T : \mathbb{R}^d \to \mathbb{R}^d, \qquad T_{\#}\rho_0 = \rho_1.$$

Density formula:

$$\rho_1(T(x)) = \rho_0(x) \left|\det \nabla T(x)\right|^{-1}.$$

Continuous normalizing flows use:

$$\dot{X}_t = b_\theta(t, X_t),$$

with likelihood:

$$\log \rho_1(X_1) = \log \rho_0(X_0) - \int_0^1 \nabla \cdot b_\theta(t, X_t)\, \mathrm{d}t.$$

Training by maximum likelihood can require differentiating through ODE solves.

Score-based diffusion prescribes a noising process:

$$\mathrm{d}X_t = f(t, X_t)\,\mathrm{d}t + g(t)\,\mathrm{d}W_t.$$

The reverse process has drift:

$$-f(1-s, x) + g^2(1-s)\nabla \log q_{1-s}(x).$$

The unknown object is the score:

$$\nabla \log q_t(x).$$

It is learned by denoising score matching:

$$\mathbb{E}\left[\lambda(t)\,\|s_\theta(t, x_t) - \nabla_{x_t} \log q(x_t \mid x_0)\|^2\right].$$

A stochastic bridge conditions a process on endpoints. For example, Brownian bridge:

$$B_0 = 0, \qquad B_1 = 0.$$

General bridge constructions often require a Doob h-transform:

$$\text{bridge drift} = \text{original drift} + a\nabla \log h.$$

The function h is generally intractable.

Stochastic interpolants avoid this by specifying directly:

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

No Doob transform is required to sample x_t during training.

Flow matching constructs conditional probability paths and learns a marginal velocity:

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

Stochastic interpolants generalize this viewpoint by allowing:

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

where both endpoints can be non-Gaussian.

They also introduce:

$$s(t, x) = \nabla \log \rho(t, x)$$

and a tunable diffusion coefficient:

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

Optimal transport asks for a map or coupling minimizing transport cost:

$$\inf_{\pi \in \Pi(\rho_0, \rho_1)} \int |x_1 - x_0|^2 \pi(\,\mathrm{d}x_0,\, \mathrm{d}x_1).$$

The dynamic Benamou–Brenier formulation:

$$\inf_{\rho, v} \int_0^1 \int |v(t,x)|^2 \rho(t,x)\,\mathrm{d}x\,\mathrm{d}t$$

subject to:

$$\partial_t \rho + \nabla \cdot (v\rho) = 0.$$

Stochastic interpolants can be optimized over to connect with optimal transport and Schrödinger bridges.

Schrödinger bridge problem:

$$\min_{\rho,u} \int_0^1 \int |u(t,x)|^2 \rho(t,x)\,\mathrm{d}x\,\mathrm{d}t$$

subject to:

$$\partial_t \rho + \nabla \cdot (u\rho) = \varepsilon \Delta \rho,$$
$$\rho(0) = \rho_0, \qquad \rho(1) = \rho_1.$$

It is entropy-regularized transport.

The stochastic interpolant paper shows how a max-min optimization over interpolants recovers the Schrödinger bridge.

A central question:

If we learn an approximate drift/score, how close is the generated law?

Distances include:

$$\mathrm{KL}, \qquad \mathrm{TV}, \qquad W_2, \qquad \text{likelihood gap.}$$

The stochastic interpolant result:

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

for stochastic dynamics.

For deterministic dynamics, drift regression alone is not enough; Fisher control appears.

Given:

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

The paper proves:

$$\rho(t) = \mathcal{L}(x_t)$$

is smooth, positive, and solves:

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

where:

$$b(t, x) = \mathbb{E}[\dot{x}_t \mid x_t = x].$$

It also proves:

$$s(t, x) = \nabla \log \rho(t, x) = -\gamma^{-1}(t)\mathbb{E}[z \mid x_t = x].$$

Both b and s are learnable by quadratic objectives.

Once b and s are known, define:

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

Then $\rho(t)$ solves:

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

and:

$$\partial_t \rho + \nabla \cdot (b^B \rho) = -\varepsilon \Delta \rho.$$

Therefore:

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

and the corresponding backward SDE sample the same path of laws.

The proofs rely on four repeated mechanisms:

1. Fourier analysis

$$g(t,k) = \mathbb{E}e^{ik\cdot x_t}.$$

2. Conditional expectation

$$\mathbb{E}[f \mid x_t = x]\rho(t,x) = \text{density-weighted conditional mean.}$$

3. Integration by parts

$$\int \phi \nabla \cdot j = -\int \nabla\phi \cdot j.$$

4. Quadratic completion

$$\frac{1}{2}|\hat{f}|^2 - f\cdot\hat{f} = \frac{1}{2}|\hat{f} - f|^2 - \frac{1}{2}|f|^2.$$

The latent Gaussian term gives:

$$\rho(t,\cdot) \in C^\infty(\mathbb{R}^d) \qquad 0 < t < 1.$$

Endpoint regularity is controlled by assumptions:

$$\rho_i > 0, \qquad \rho_i \in C^2(\mathbb{R}^d), \qquad \int |\nabla \log \rho_i|^2 \rho_i < \infty.$$

The interpolant regularity assumptions control:

$$\mathbb{E}|\partial_t I|^4, \qquad \mathbb{E}|\partial_t^2 I|^2.$$

These ensure $b \in L^2(\rho)$ and the objectives are finite.

Although $\gamma(0) = \gamma(1) = 0$, the score formula contains:

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

Thus $s(t, x)$ is naturally defined for:

$$0 < t < 1.$$

Endpoint statements require limits and Fisher information assumptions.

When:

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

then:

$$\lim_{t\to 0} \gamma(t)\dot{\gamma}(t) = \frac{a}{2}, \qquad \lim_{t\to 1} \gamma(t)\dot{\gamma}(t) = -\frac{a}{2}.$$

Endpoint velocities can contain endpoint scores.

By the end of Semester I, students should prove:

$$x_t = I + \gamma z \quad \Rightarrow \quad \rho(t) > 0,\ \rho(t) \in C^\infty.$$

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

$$b = \arg\min L_b, \qquad s = \arg\min L_s, \qquad \eta_z = \arg\min L_{\eta_z}.$$

$$b^F = b + \varepsilon s \quad \Rightarrow \quad \partial_t \rho + \nabla \cdot (b^F \rho) = \varepsilon \Delta \rho.$$

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

SUGGESTED WEEKLY STRUCTURE

Weeks	Focus
1	Measure, weak PDE, conditional expectation
2	Fourier analysis, Gaussian smoothing
3	Transport equations and CNF likelihood
4	Itô calculus and Fokker–Planck
5	Stochastic interpolant definition and Theorem 6
6	Quadratic objectives and score identity
7	ODE/SDE generative models
8	KL control and likelihood theory
9	Density estimation and cross-entropy
10	Diffusive and one-sided interpolants
11	Schrödinger bridges
12	Spatially linear interpolants
13	Connections and algorithms
14	Student proof presentations

The rest of the semester depends on the following dictionary:

$$
\begin{aligned}
\text{interpolant} &: \quad x_t = I(t, x_0, x_1) + \gamma(t) z, \\
\text{density path} &: \quad \rho(t) = \mathcal{L}(x_t), \\
\text{velocity} &: \quad b(t, x) = \mathbb{E}[\dot{x}_t \mid x_t = x], \\
\text{score} &: \quad s(t, x) = \nabla \log \rho(t, x), \\
\text{transport PDE} &: \quad \partial_t \rho + \nabla \cdot (b\rho) = 0, \\
\text{Fokker–Planck PDE} &: \quad \partial_t \rho + \nabla \cdot ((b + \varepsilon s)\rho) = \varepsilon \Delta \rho.
\end{aligned}
$$

Next document:

Definitions, assumptions, transport equation, score identity, and quadratic objectives.