Lecture Notes: From Particle Dynamics and Transport PDEs to Generative AI

A Rigorous 90-Minute Bridge via Optimal Transport, Fokker–Planck, and Score Models

1 Learning goals and roadmap

These notes give a mathematically rigorous bridge from deterministic transport and stochastic dynamics to modern generative AI.

Learning goals. By the end of the lecture, students should be able to:

(1) Prove the link between a particle flow ODE and the continuity equation for the density.

(2) Derive the Fokker–Planck equation from an Itô SDE in weak form.

(3) Understand the dynamic optimal transport formulation (Benamou–Brenier) in a smooth setting.

(4) See why the score $\nabla \log \rho_t$ naturally appears in diffusion-based generative models.

(5) Interpret sampling in generative AI as numerical integration of dynamics in probability space.

Lecture roadmap (90 minutes).

(1) Recap: particle ODE and transport PDE (10 min)

(2) Deterministic transport and continuity equation, with proof (20 min)

(3) Dynamic OT bridge (Benamou–Brenier) in the smooth setting, with proof (20 min)

(4) Stochastic transport and Fokker–Planck, with proof (20 min)

(5) Probability current, score, and probability flow ODE (15 min)

(6) Generative AI interpretation and numerical sampling (5 min)

2 Notation and assumptions

We work on $\mathbb{R}^d$ throughout. Time is $t \in [0, T]$ unless stated otherwise.

(1) A time-dependent vector field is denoted by $v_t(x)$ or $b_t(x)$.

(2) A density is denoted by $\rho_t(x)$, always assumed nonnegative with $\int_{\mathbb{R}^d} \rho_t(x)\,\mathrm{d}x = 1$.

(3) The spatial gradient is ∇, divergence is $\nabla\cdot$, and Laplacian is Δ.

(4) Pushforward of a measure μ under a map Φ is denoted by $\Phi_{\#}\mu$.

(5) For an SDE $dX_t = b_t(X_t)\,\mathrm{d}t + \sigma_t(X_t)\,\mathrm{d}W_t$, define

$$a_t(x) := \sigma_t(x)\sigma_t(x)^\top \in \mathbb{R}^{d\times d}.$$

(6) All proofs are written in a smooth setting to keep the derivations transparent. The same identities extend to weaker settings under standard assumptions.

3 Recap: particle dynamics and density evolution

We start from the deterministic particle ODE

$$\dot{X}_t = v_t(X_t), \qquad X_0 \sim \rho_0. \tag{1}$$

If the random initial condition X_0 has density ρ_0, then the law of X_t is a density ρ_t, and the pair (ρ_t, v_t) is expected to satisfy the continuity equation

$$\partial_t \rho_t + \nabla\cdot(\rho_t v_t) = 0. \tag{2}$$

Guiding question for generative AI. How should we choose a time-dependent field v_t so that a simple initial density (such as a Gaussian) evolves into a target data density?

This is the transport viewpoint of generative modeling.

4 Deterministic transport: flow map and continuity equation

4.1 Flow map and pushforward

Let $v \in C^1([0,T]\times\mathbb{R}^d;\mathbb{R}^d)$. Denote by $\Phi_t(x)$ the flow map solving

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t(x) = v_t(\Phi_t(x)), \qquad \Phi_0(x) = x. \tag{3}$$

If $X_0 \sim \rho_0$, then $X_t = \Phi_t(X_0)$ and the law of X_t is

$$\rho_t = (\Phi_t)_{\#}\rho_0.$$

4.2 Main theorem: continuity equation in weak form

Theorem 4.1 (Flow map implies continuity equation)**.** *Assume $v \in C^1([0,T]\times\mathbb{R}^d;\mathbb{R}^d)$ and let Φ_t be the associated flow map (3). Let $\rho_0 \in C_c^1(\mathbb{R}^d)$ be a probability density, and define $\rho_t = (\Phi_t)_{\#}\rho_0$.*

Then, for every test function $\varphi \in C_c^\infty(\mathbb{R}^d)$, one has

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d}\varphi(x)\rho_t(x)\,\mathrm{d}x = \int_{\mathbb{R}^d}\nabla\varphi(x)\cdot v_t(x)\rho_t(x)\,\mathrm{d}x. \tag{4}$$

Equivalently, ρ_t solves the continuity equation (2) in the weak sense.

Proof. By the definition of pushforward,

$$\int_{\mathbb{R}^d}\varphi(x)\rho_t(x)\,\mathrm{d}x = \int_{\mathbb{R}^d}\varphi(\Phi_t(x))\rho_0(x)\,\mathrm{d}x.$$

Since v is C^1 and φ is smooth with compact support, differentiation under the integral is justified:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d}\varphi(\Phi_t(x))\rho_0(x)\,\mathrm{d}x = \int_{\mathbb{R}^d}\nabla\varphi(\Phi_t(x))\cdot\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t(x)\,\rho_0(x)\,\mathrm{d}x.$$

Using the ODE (3),

$$= \int_{\mathbb{R}^d}\nabla\varphi(\Phi_t(x))\cdot v_t(\Phi_t(x))\,\rho_0(x)\,\mathrm{d}x.$$

Apply the pushforward identity again with $y = \Phi_t(x)$:

$$= \int_{\mathbb{R}^d}\nabla\varphi(y)\cdot v_t(y)\rho_t(y)\,\mathrm{d}y,$$

which proves (4).

To recover the PDE form, integrate by parts:

$$\int_{\mathbb{R}^d}\nabla\varphi\cdot(v_t\rho_t)\,\mathrm{d}x = -\int_{\mathbb{R}^d}\varphi\,\nabla\cdot(\rho_t v_t)\,\mathrm{d}x.$$

Hence

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d}\varphi\rho_t\,\mathrm{d}x = -\int_{\mathbb{R}^d}\varphi\,\nabla\cdot(\rho_t v_t)\,\mathrm{d}x,$$

which is exactly the weak form of (2). This completes the proof. □

Remark 4.2 (Strong form via Jacobian identity)**.** If each Φ_t is a C^1 diffeomorphism, then one has

$$\rho_t(\Phi_t(x))\det(D\Phi_t(x)) = \rho_0(x).$$

Differentiating this identity in time also yields $\partial_t\rho_t + \nabla\cdot(\rho_t v_t) = 0$ in strong form.

4.3 Generative AI interpretation

A *flow-based generative model* parameterizes a velocity field $v_\theta(t,x)$ using a neural network, then samples by solving

$$\dot{X}_t = v_\theta(t, X_t), \qquad X_0 \sim \rho_{\text{prior}}.$$

By Theorem 4.1, the resulting density evolves through a continuity equation. In other words, the learned particle dynamics induces a learned transport PDE on the density.

5 Optional bridge: dynamic optimal transport (Benamou–Brenier) with proof

This section provides the rigorous bridge from the continuity equation to optimal transport in a smooth setting.

5.1 Statement

Theorem 5.1 (Benamou–Brenier formula in a smooth setting)**.** *Let ρ_0, ρ_1 be probability densities on $\mathbb{R}^d$ with finite second moments. Assume:*

(i) For every admissible pair (ρ_t, v_t), the velocity field is smooth enough so that the ODE flow Φ_t is a global diffeomorphism on $\mathbb{R}^d$, and $\rho_t = (\Phi_t)_\#\rho_0$.

(ii) There exists an optimal transport map $T : \mathbb{R}^d \to \mathbb{R}^d$ for the quadratic cost, pushing ρ_0 to ρ_1, and the displacement interpolation

$$\Phi_t(x) = (1-t)x + tT(x)$$

is a diffeomorphism for each $t \in [0,1]$.

Then

$$W_2^2(\rho_0,\rho_1) = \inf_{\substack{(\rho_t,v_t)\\ \partial_t\rho_t+\nabla\cdot(\rho_t v_t)=0\\ \rho_0=\rho_0,\ \rho_1=\rho_1}} \int_0^1 \int_{\mathbb{R}^d} |v_t(x)|^2 \rho_t(x)\,\mathrm{d}x\,\mathrm{d}t. \tag{5}$$

Moreover, the infimum is attained by the displacement interpolation generated by T.

5.2 Proof

Proof. We prove the lower bound and upper bound separately.

Step 1: Lower bound. Let (ρ_t, v_t) be any admissible smooth pair in (5). By assumption, there exists a smooth flow Φ_t such that

$$\dot\Phi_t(x) = v_t(\Phi_t(x)), \qquad \Phi_0(x) = x, \qquad \rho_t = (\Phi_t)_\#\rho_0.$$

Since $\rho_1 = (\Phi_1)_\#\rho_0$, the map Φ_1 transports ρ_0 to ρ_1. Therefore the coupling

$$\pi := (\mathrm{Id}, \Phi_1)_\#\rho_0$$

is admissible in the definition of W_2. Hence

$$W_2^2(\rho_0,\rho_1) \le \int_{\mathbb{R}^d} |x - \Phi_1(x)|^2 \rho_0(x)\,\mathrm{d}x.$$

Now write

$$\Phi_1(x) - x = \int_0^1 \dot\Phi_t(x)\,\mathrm{d}t = \int_0^1 v_t(\Phi_t(x))\,\mathrm{d}t.$$

By Jensen's inequality in time,

$$\left|\int_0^1 v_t(\Phi_t(x))\,\mathrm{d}t\right|^2 \le \int_0^1 |v_t(\Phi_t(x))|^2\,\mathrm{d}t.$$

Integrating against $\rho_0(x)\,\mathrm{d}x$ gives

$$\int_{\mathbb{R}^d} |x - \Phi_1(x)|^2 \rho_0(x)\,\mathrm{d}x \le \int_{\mathbb{R}^d}\int_0^1 |v_t(\Phi_t(x))|^2\,\mathrm{d}t\,\rho_0(x)\,\mathrm{d}x.$$

Use Fubini and the pushforward identity $\rho_t = (\Phi_t)_\#\rho_0$:

$$\int_{\mathbb{R}^d}\int_0^1 |v_t(\Phi_t(x))|^2\,\mathrm{d}t\,\rho_0(x)\,\mathrm{d}x = \int_0^1\int_{\mathbb{R}^d} |v_t(y)|^2\rho_t(y)\,\mathrm{d}y\,\mathrm{d}t.$$

Combining these estimates, we obtain

$$W_2^2(\rho_0,\rho_1) \le \int_0^1 \int_{\mathbb{R}^d} |v_t(y)|^2\rho_t(y)\,\mathrm{d}y\,\mathrm{d}t.$$

Since (ρ_t, v_t) was arbitrary admissible, this implies

$$W_2^2(\rho_0,\rho_1) \le \inf_{\text{admissible}} \int_0^1 \int_{\mathbb{R}^d} |v_t|^2\rho_t. \tag{6}$$

Step 2: Upper bound via displacement interpolation. Let T be an optimal quadratic transport map from ρ_0 to ρ_1. Define

$$\Phi_t(x) = (1-t)x + tT(x), \qquad \rho_t = (\Phi_t)_{\#}\rho_0.$$

Then $\Phi_0 = \mathrm{Id}$ and $\Phi_1 = T$, so ρ_0 and ρ_1 are the endpoints.

Define the Eulerian velocity field v_t by

$$v_t(\Phi_t(x)) := T(x) - x.$$

This is well-defined and smooth by the diffeomorphism assumption on Φ_t. Since

$$\dot{\Phi}_t(x) = T(x) - x = v_t(\Phi_t(x)),$$

Theorem 4.1 implies that (ρ_t, v_t) satisfies the continuity equation.

Its kinetic action is

$$\int_0^1 \int_{\mathbb{R}^d} |v_t(y)|^2 \rho_t(y)\,\mathrm{d}y\,\mathrm{d}t = \int_0^1 \int_{\mathbb{R}^d} |v_t(\Phi_t(x))|^2 \rho_0(x)\,\mathrm{d}x\,\mathrm{d}t$$

$$= \int_0^1 \int_{\mathbb{R}^d} |T(x) - x|^2 \rho_0(x)\,\mathrm{d}x\,\mathrm{d}t = \int_{\mathbb{R}^d} |T(x) - x|^2 \rho_0(x)\,\mathrm{d}x.$$

Since T is optimal for the quadratic cost,

$$\int_{\mathbb{R}^d} |T(x) - x|^2 \rho_0(x)\,\mathrm{d}x = W_2^2(\rho_0, \rho_1).$$

Therefore

$$\inf_{\text{admissible}} \int_0^1 \int_{\mathbb{R}^d} |v_t|^2 \rho_t \le W_2^2(\rho_0, \rho_1). \tag{7}$$

Combining (6) and (7) yields (5). □

Remark 5.2 (Interpretation for generative modeling)**.** The Benamou–Brenier formula shows that transporting a prior density ρ_0 to a data density ρ_1 can be viewed as finding a path in probability space with minimal kinetic energy. This is a precise mathematical bridge from transport PDEs to flow-based generative models.

6 Stochastic transport: SDE and Fokker–Planck equation

We now add noise to the particle dynamics. The microscopic model becomes an Itô SDE

$$\mathrm{d}X_t = b_t(X_t)\,\mathrm{d}t + \sigma_t(X_t)\,\mathrm{d}W_t, \tag{8}$$

where W_t is a standard m-dimensional Brownian motion (often $m = d$).

6.1 Generator and weak form

For a smooth test function $\varphi : \mathbb{R}^d \to \mathbb{R}$, define the (time-dependent) generator

$$(\mathcal{L}_t\varphi)(x) = b_t(x) \cdot \nabla\varphi(x) + \frac{1}{2}\mathrm{Tr}(a_t(x)\nabla^2\varphi(x)), \qquad a_t(x) = \sigma_t(x)\sigma_t(x)^\top. \tag{9}$$

6.2 Main theorem: Fokker–Planck in weak form

Theorem 6.1 (SDE implies Fokker–Planck equation)**.** *Assume b, σ are smooth enough and satisfy standard growth and Lipschitz conditions so that (8) has a unique nonexplosive solution. Let ρ_t be the density of X_t.*

Then for every $\varphi \in C_c^\infty(\mathbb{R}^d)$,

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathbb{R}^d} \varphi(x)\rho_t(x)\,\mathrm{d}x = \int_{\mathbb{R}^d} (\mathcal{L}_t\varphi)(x)\rho_t(x)\,\mathrm{d}x. \tag{10}$$

Equivalently, ρ_t solves the Fokker–Planck equation

$$\partial_t \rho_t = -\nabla\cdot(b_t\rho_t) + \frac{1}{2}\sum_{i,j=1}^{d} \partial_{ij}(a_{ij,t}\rho_t) \tag{11}$$

in the weak sense.

Proof. Let $\varphi \in C_c^\infty(\mathbb{R}^d)$. By Itô's formula applied to $\varphi(X_t)$,

$$\mathrm{d}\varphi(X_t) = \nabla\varphi(X_t)\cdot b_t(X_t)\,\mathrm{d}t + \frac{1}{2}\mathrm{Tr}\big(a_t(X_t)\nabla^2\varphi(X_t)\big)\,\mathrm{d}t + \nabla\varphi(X_t)\cdot\sigma_t(X_t)\,\mathrm{d}W_t.$$

Taking expectation and using that the stochastic integral has mean zero,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[\varphi(X_t)] = \mathbb{E}\left[\nabla\varphi(X_t)\cdot b_t(X_t) + \frac{1}{2}\mathrm{Tr}\big(a_t(X_t)\nabla^2\varphi(X_t)\big)\right].$$

Since X_t has density ρ_t,

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d}\varphi(x)\rho_t(x)\,\mathrm{d}x = \int_{\mathbb{R}^d}\left[b_t(x)\cdot\nabla\varphi(x) + \frac{1}{2}\mathrm{Tr}(a_t(x)\nabla^2\varphi(x))\right]\rho_t(x)\,\mathrm{d}x.$$

This is (10).

To identify the PDE form, integrate by parts. For the drift term,

$$\int_{\mathbb{R}^d} b_t\cdot\nabla\varphi\,\rho_t\,\mathrm{d}x = -\int_{\mathbb{R}^d}\varphi\,\nabla\cdot(b_t\rho_t)\,\mathrm{d}x.$$

For the second-order term, write

$$\mathrm{Tr}(a_t\nabla^2\varphi) = \sum_{i,j=1}^{d} a_{ij,t}\partial_{ij}\varphi.$$

By two integrations by parts,

$$\int_{\mathbb{R}^d} a_{ij,t}\partial_{ij}\varphi\,\rho_t\,\mathrm{d}x = \int_{\mathbb{R}^d}\varphi\,\partial_{ij}(a_{ij,t}\rho_t)\,\mathrm{d}x.$$

Summing over i, j yields

$$\int_{\mathbb{R}^d}\mathrm{Tr}(a_t\nabla^2\varphi)\rho_t\,\mathrm{d}x = \int_{\mathbb{R}^d}\varphi\sum_{i,j}\partial_{ij}(a_{ij,t}\rho_t)\,\mathrm{d}x.$$

Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d}\varphi\rho_t\,\mathrm{d}x = \int_{\mathbb{R}^d}\varphi\left[-\nabla\cdot(b_t\rho_t) + \frac{1}{2}\sum_{i,j}\partial_{ij}(a_{ij,t}\rho_t)\right]\mathrm{d}x.$$

This is the weak form of (11). □

7 Lecture Module: Multidimensional Itô Formula (Classroom Version)

7.1 Learning objectives

By the end of this module, students should be able to:

1. State the multidimensional Itô formula.
2. Understand why a second-order term appears (quadratic variation).
3. Use Itô formula to identify the generator
$$(\mathcal{L}_t f)(x) = b_t(x) \cdot \nabla f(x) + \frac{1}{2}\mathrm{Tr}\big(a_t(x)\nabla^2 f(x)\big), \qquad a_t(x) = \sigma_t(x)\sigma_t(x)^\top.$$
4. Apply Itô formula as preparation for the Fokker–Planck derivation.

7.2 Setup and notation

Let $W_t \in \mathbb{R}^m$ be an m-dimensional Brownian motion, and let $X_t \in \mathbb{R}^d$ solve the SDE
$$dX_t = b_t\,dt + \sigma_t\,dW_t, \tag{12}$$
where
$$b_t \in \mathbb{R}^d, \qquad \sigma_t \in \mathbb{R}^{d\times m}.$$
Assume b_t, σ_t are sufficiently regular (adapted, integrable, and smooth enough for the manipulations below).

Let $f \in C^2(\mathbb{R}^d)$.

7.3 Main theorem

Theorem 7.1 (Multidimensional Itô formula)**.** *Let X_t solve (12). Then for every $f \in C^2(\mathbb{R}^d)$,*
$$f(X_t) - f(X_0) = \int_0^t \nabla f(X_s)\cdot b_s\,ds + \int_0^t \nabla f(X_s)\cdot \sigma_s\,dW_s + \frac{1}{2}\int_0^t \mathrm{Tr}\big(\sigma_s\sigma_s^\top \nabla^2 f(X_s)\big)\,ds. \tag{13}$$
Equivalently, in differential form,
$$df(X_t) = \nabla f(X_t)\cdot b_t\,dt + \nabla f(X_t)\cdot \sigma_t\,dW_t + \frac{1}{2}\,\mathrm{Tr}\big(\sigma_t\sigma_t^\top \nabla^2 f(X_t)\big)\,dt. \tag{14}$$

7.4 Proof idea for class (Taylor expansion + quadratic variation)

Classroom proof sketch. We give a proof that highlights the main mechanism.

Step 1: Partition and telescoping sum. Take a partition
$$0 = t_0 < t_1 < \cdots < t_n = t,$$
and write
$$f(X_t) - f(X_0) = \sum_{i=0}^{n-1} \big(f(X_{t_{i+1}}) - f(X_{t_i})\big).$$
Set
$$\Delta_i X := X_{t_{i+1}} - X_{t_i}, \qquad \Delta_i t := t_{i+1} - t_i, \qquad \Delta_i W := W_{t_{i+1}} - W_{t_i}.$$

Step 2: Second-order Taylor expansion. Expand f at X_{t_i}:

$$f(X_{t_{i+1}}) - f(X_{t_i}) = \nabla f(X_{t_i}) \cdot \Delta_i X + \frac{1}{2} \Delta_i X^\top \nabla^2 f(X_{t_i}) \Delta_i X + r_i,$$

where the remainder $r_i = o(|\Delta_i X|^2)$.

Step 3: Approximate the increment $\Delta_i X$. From the SDE,

$$\Delta_i X = \int_{t_i}^{t_{i+1}} b_s \, ds + \int_{t_i}^{t_{i+1}} \sigma_s \, dW_s.$$

On a small interval, the leading-order approximation is

$$\Delta_i X \approx b_{t_i} \Delta_i t + \sigma_{t_i} \Delta_i W.$$

Step 4: First-order term converges to drift + stochastic integral. The first-order sum becomes

$$\sum_i \nabla f(X_{t_i}) \cdot \Delta_i X \approx \sum_i \nabla f(X_{t_i}) \cdot b_{t_i} \Delta_i t + \sum_i \nabla f(X_{t_i}) \cdot \sigma_{t_i} \Delta_i W.$$

As the mesh size goes to zero, these converge to

$$\int_0^t \nabla f(X_s) \cdot b_s \, ds \quad \text{and} \quad \int_0^t \nabla f(X_s) \cdot \sigma_s \, dW_s.$$

Step 5: Second-order term and quadratic variation. Consider

$$\frac{1}{2} \sum_i \Delta_i X^\top \nabla^2 f(X_{t_i}) \Delta_i X.$$

Substitute the increment approximation:

$$\Delta_i X \approx b_{t_i} \Delta_i t + \sigma_{t_i} \Delta_i W.$$

When expanded, there are three types of contributions:

1. $(b\Delta t)^\top H (b\Delta t)$, size $O((\Delta t)^2)$,
2. cross terms $(b\Delta t)^\top H (\sigma \Delta W)$, size $O((\Delta t)^{3/2})$,
3. $(\sigma \Delta W)^\top H (\sigma \Delta W)$, size $O(\Delta t)$.

Only the third type survives after summation.

Thus the leading second-order contribution is

$$\frac{1}{2} \sum_i (\sigma_{t_i} \Delta_i W)^\top \nabla^2 f(X_{t_i}) (\sigma_{t_i} \Delta_i W).$$

Rewrite it using the trace identity $u^\top H u = \mathrm{Tr}(H u u^\top)$:

$$(\sigma_{t_i} \Delta_i W)^\top \nabla^2 f(X_{t_i}) (\sigma_{t_i} \Delta_i W) = \mathrm{Tr}\Big(\nabla^2 f(X_{t_i}) \, \sigma_{t_i} (\Delta_i W \Delta_i W^\top) \sigma_{t_i}^\top \Big).$$

Now use the quadratic variation of Brownian motion:

$$\Delta_i W \Delta_i W^\top \approx I_m \, \Delta_i t \quad \text{in the summed limit.}$$

Therefore,

$$\frac{1}{2} \sum_i \Delta_i X^\top \nabla^2 f(X_{t_i}) \Delta_i X \to \frac{1}{2} \int_0^t \mathrm{Tr}\big(\sigma_s \sigma_s^\top \nabla^2 f(X_s)\big) \, ds.$$

Step 6: Remainder term vanishes. Under standard regularity assumptions, the Taylor remainders satisfy

$$\sum_i r_i \to 0$$

(in probability, or in L^1, depending on assumptions).

Combining Steps 4–6 gives (13). □

7.5 Symbolic differential rule (for intuition)

A compact way to remember the formula is:

$$df(X_t) = \nabla f(X_t) \cdot dX_t + \frac{1}{2}\, dX_t^\top \nabla^2 f(X_t)\, dX_t,$$

together with the Itô multiplication rules

$$dt\, dt = 0, \qquad dt\, dW_t = 0, \qquad dW_t\, dW_t^\top = I_m\, dt.$$

Since $dX_t = b_t\, dt + \sigma_t\, dW_t$, one gets

$$dX_t\, dX_t^\top = \sigma_t (dW_t\, dW_t^\top) \sigma_t^\top = \sigma_t \sigma_t^\top\, dt,$$

which yields exactly (14).

7.6 Generator form (used later in Fokker–Planck)

Define

$$a_t(x) := \sigma_t(x)\sigma_t(x)^\top.$$

Then Itô formula can be written as

$$df(X_t) = (\mathcal{L}_t f)(X_t)\, dt + \nabla f(X_t) \cdot \sigma_t(X_t)\, dW_t,$$

where

$$(\mathcal{L}_t f)(x) = b_t(x) \cdot \nabla f(x) + \frac{1}{2} \mathrm{Tr}\big(a_t(x) \nabla^2 f(x)\big). \tag{15}$$

This is the infinitesimal generator that appears in the weak derivation of the Fokker–Planck equation.

7.7 Example for class

Take $d = m = 1$, $X_t = W_t$, and $f(x) = x^2$. Then

$$df(W_t) = 2W_t\, dW_t + \frac{1}{2} \cdot 2\, dt = 2W_t\, dW_t + dt.$$

Integrating,

$$W_t^2 = t + 2\int_0^t W_s\, dW_s.$$

This is the simplest illustration of why the extra second-order term appears.

7.8 Suggested in-class timing (20 minutes)

1. Statement and notation (3 min)
2. Taylor expansion proof idea (8 min)
3. Quadratic variation explanation and trace form (6 min)
4. Generator form + one example (3 min)

7.9 Optional exercise

Let $X_t \in \mathbb{R}^d$ solve

$$dX_t = b(X_t)\,dt + \sigma(X_t)\,dW_t.$$

Apply Itô formula to $f(x) = |x|^2$ and show

$$d|X_t|^2 = 2X_t \cdot b(X_t)\,dt + 2X_t \cdot \sigma(X_t)\,dW_t + \mathrm{Tr}\big(\sigma(X_t)\sigma(X_t)^\top\big)\,dt.$$

Remark 7.2 (Isotropic noise case)**.** If $\sigma_t(x) = \sqrt{2\beta_t}\,I_d$ with scalar $\beta_t \geq 0$, then $a_t(x) = 2\beta_t I_d$ and (11) simplifies to

$$\partial_t \rho_t = -\nabla\cdot(b_t\rho_t) + \beta_t \Delta\rho_t. \tag{16}$$

7.10 Generative AI interpretation

Diffusion-based generative models use a *forward noising process* that maps the data distribution into a simple prior (often approximately Gaussian). The learned reverse dynamics then reconstructs samples from the prior to the data distribution. The Fokker–Planck equation (11) is the macroscopic evolution equation behind this construction.

8 Probability current, score, and the probability flow ODE

This section explains why score functions appear naturally.

8.1 Probability current form of Fokker–Planck

For simplicity, consider the isotropic diffusion case

$$\mathrm{d}X_t = f_t(X_t)\,\mathrm{d}t + g_t\,\mathrm{d}W_t, \tag{17}$$

where $g_t \geq 0$ is scalar (time-dependent only), and let ρ_t be its density. Then (16) becomes

$$\partial_t \rho_t = -\nabla\cdot(f_t\rho_t) + \frac{g_t^2}{2}\Delta\rho_t. \tag{18}$$

Define the *probability current*

$$J_t := f_t\rho_t - \frac{g_t^2}{2}\nabla\rho_t. \tag{19}$$

Then (18) can be written as

$$\partial_t \rho_t + \nabla\cdot J_t = 0. \tag{20}$$

8.2 Score and equivalent transport form

Whenever $\rho_t(x) > 0$, define the *score function*

$$s_t(x) := \nabla \log \rho_t(x). \tag{21}$$

Since $\nabla \rho_t = \rho_t \nabla \log \rho_t$, (19) becomes

$$J_t = \left(f_t - \frac{g_t^2}{2} s_t \right) \rho_t.$$

Hence (20) is equivalent to the continuity equation

$$\partial_t \rho_t + \nabla \cdot (\rho_t v_t^{\mathrm{pf}}) = 0 \tag{22}$$

with velocity field

$$v_t^{\mathrm{pf}}(x) := f_t(x) - \frac{g_t^2}{2} s_t(x) = f_t(x) - \frac{g_t^2}{2} \nabla \log \rho_t(x). \tag{23}$$

This motivates the *probability flow ODE*

$$\dot{X}_t = v_t^{\mathrm{pf}}(X_t) = f_t(X_t) - \frac{g_t^2}{2} \nabla \log \rho_t(X_t). \tag{24}$$

Proposition 8.1 (Equivalent density evolution)**.** *Assume ρ_t is a smooth positive solution of (18). Then the same density ρ_t also satisfies the continuity equation (22) with velocity (23).*

Proof. Starting from (18),

$$\partial_t \rho_t = -\nabla \cdot (f_t \rho_t) + \frac{g_t^2}{2} \Delta \rho_t.$$

Since g_t depends only on time,

$$\frac{g_t^2}{2} \Delta \rho_t = \nabla \cdot \left(\frac{g_t^2}{2} \nabla \rho_t \right).$$

Thus

$$\partial_t \rho_t = -\nabla \cdot \left(f_t \rho_t - \frac{g_t^2}{2} \nabla \rho_t \right).$$

Using $\nabla \rho_t = \rho_t \nabla \log \rho_t$,

$$f_t \rho_t - \frac{g_t^2}{2} \nabla \rho_t = \left(f_t - \frac{g_t^2}{2} \nabla \log \rho_t \right) \rho_t = v_t^{\mathrm{pf}} \rho_t.$$

Hence

$$\partial_t \rho_t = -\nabla \cdot (v_t^{\mathrm{pf}} \rho_t),$$

which is exactly (22). □

Remark 8.2 (Why the score appears)**.** The score is not an arbitrary quantity introduced by machine learning practice. It appears because it is the logarithmic gradient needed to rewrite the Fokker–Planck PDE as a transport PDE with a deterministic velocity field.

9 Connection to diffusion models and score-based generative AI

9.1 Forward noising process

In diffusion models, the forward process gradually transforms data into noise. A common formulation is

$$\mathrm{d}X_t = f_t(X_t)\,\mathrm{d}t + g_t\,\mathrm{d}W_t, \qquad X_0 \sim \rho_{\text{data}}. \tag{25}$$

Its density evolves according to (18).

9.2 Reverse-time generation and score estimation

The reverse dynamics depends on the score $s_t(x) = \nabla \log \rho_t(x)$. Since ρ_t is unknown, one trains a neural network $s_\theta(t, x)$ to approximate $s_t(x)$.

A standard practical route is denoising score matching. For example, in a Gaussian perturbation setting,

$$X_t = \alpha_t X_0 + \sigma_t Z, \qquad Z \sim \mathcal{N}(0, I_d),$$

one can train a network to predict a quantity equivalent to the score (or the noise) using supervised losses generated from noisy data pairs.

Remark 9.1 (Pedagogical point)**.** At a conceptual level, diffusion model training is *not* direct density estimation. It is estimation of a family of vector fields $s_t(\cdot)$ indexed by time, which encode the local geometry of the evolving distribution.

9.3 Flow models versus diffusion models

(1) **Flow model:** deterministic particle dynamics

$$\dot{X}_t = v_\theta(t, X_t), \qquad \partial_t \rho_t + \nabla \cdot (\rho_t v_\theta) = 0.$$

(2) **Diffusion model:** stochastic particle dynamics

$$\mathrm{d}X_t = f_t(X_t)\,\mathrm{d}t + g_t\,\mathrm{d}W_t, \qquad \partial_t \rho_t = -\nabla \cdot (f_t \rho_t) + \tfrac{g_t^2}{2} \Delta \rho_t.$$

(3) **Probability flow view:** deterministic reformulation of the density evolution

$$\dot{X}_t = f_t(X_t) - \tfrac{g_t^2}{2} \nabla \log \rho_t(X_t).$$

10 Numerical sampling as computation in probability space

This final section connects directly to a course on probability-space computation.

10.1 Sampling is numerical integration

Once a generative model is trained, sample generation is a numerical integration task:

(1) **SDE sampling:** integrate an SDE using schemes such as Euler–Maruyama or predictor–corrector methods.

(2) **ODE sampling:** integrate the probability flow ODE using standard ODE solvers.

10.2 Inference-time control and conditioning

Conditioning (text prompts, observations, constraints, posterior information) can be interpreted as modifying the drift or score field. This is closely related to ideas from Bayesian inference, control, and data assimilation.

Remark 10.1 (Bridge to advanced topics)**.** This viewpoint is particularly useful in scientific machine learning, where one often wants to impose observation consistency or posterior constraints at inference time after a generative prior has already been trained.

11 Suggested board plan for a 90-minute lecture

Board 1: Recap the particle ODE and continuity equation:

$$\dot{X}_t = v_t(X_t), \qquad \partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0.$$

State the generative AI question.

Board 2: Define the flow map Φ_t, pushforward $\rho_t = (\Phi_t)_{\#}\rho_0$, and prove Theorem 4.1.

Board 3: State and prove Theorem 5.1 (Benamou–Brenier bridge, smooth setting).

Board 4: Introduce the SDE, generator $\mathcal{L}_t$, and prove Theorem 6.1.

Board 5: Write the isotropic Fokker–Planck equation, define the current J_t, define the score s_t, and derive the probability flow ODE.

Board 6: Summarize the flow-model and diffusion-model interpretations, then explain sampling as numerical integration.

12 Exercises (for homework or discussion)

(1) **Jacobian derivation of the continuity equation.** Assume Φ_t is a C^1 diffeomorphism and prove

$$\rho_t(\Phi_t(x)) \det(D\Phi_t(x)) = \rho_0(x).$$

Differentiate in time to recover $\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0$.

(2) **Heat equation as a special case.** Let

$$\mathrm{d}X_t = \sqrt{2\beta}\, \mathrm{d}W_t, \qquad \beta > 0.$$

Show that the density solves

$$\partial_t \rho_t = \beta \Delta \rho_t.$$

(3) **Probability current form.** Starting from

$$\partial_t \rho_t = -\nabla \cdot (f_t \rho_t) + \frac{g_t^2}{2} \Delta \rho_t,$$

verify directly that

$$J_t = f_t \rho_t - \frac{g_t^2}{2} \nabla \rho_t$$

gives $\partial_t \rho_t + \nabla \cdot J_t = 0$.

(4) **Benamou–Brenier lower bound.** Reproduce Step 1 in the proof of Theorem 5.1, paying attention to the use of Jensen's inequality and pushforward measures.

(5) **Modeling question.** In your own words, explain why score estimation is enough to define a generative sampling dynamics.

13 Optional extension: JKO viewpoint (one-slide mention)

If time permits, mention the Wasserstein gradient flow interpretation:

$$\rho^{k+1} \in \arg\min_{\rho \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ \mathcal{F}(\rho) + \frac{1}{2\tau} W_2^2(\rho, \rho^k) \right\}.$$

This gives a variational time-discretization of certain diffusion PDEs and provides another conceptual bridge from optimal transport to generative dynamics.

Theorem 13.1 (JKO step as particle transport: Euler–Lagrange form)**.** *Let $\rho^k \in \mathcal{P}_2(\mathbb{R}^d)$ be given. Fix $\tau > 0$, and let*

$$\rho^{k+1} \in \arg\min_{\rho \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ \mathcal{F}(\rho) + \frac{1}{2\tau} W_2^2(\rho, \rho^k) \right\}.$$

Assume the following smoothness/regularity conditions hold:

1. *ρ^{k+1} is absolutely continuous with respect to Lebesgue measure, with smooth positive density (still denoted ρ^{k+1}).*
2. *$\mathcal{F}$ admits a smooth first variation at ρ^{k+1}, i.e., there exists a smooth function*

$$\frac{\delta \mathcal{F}}{\delta \rho}(\rho^{k+1})(x)$$

such that for every smooth compactly supported vector field ξ and the perturbation

$$\rho_\varepsilon := (\mathrm{Id} + \varepsilon \xi)_\# \rho^{k+1},$$

one has

$$\left. \frac{d}{d\varepsilon} \right|_{\varepsilon=0} \mathcal{F}(\rho_\varepsilon) = \int_{\mathbb{R}^d} \nabla \left(\frac{\delta \mathcal{F}}{\delta \rho}(\rho^{k+1})(x) \right) \cdot \xi(x)\, \rho^{k+1}(x)\, dx.$$

3. *There exists a unique optimal transport map $T : \mathbb{R}^d \to \mathbb{R}^d$ from ρ^{k+1} to ρ^k for the quadratic cost:*

$$T_\# \rho^{k+1} = \rho^k, \qquad W_2^2(\rho^{k+1}, \rho^k) = \int_{\mathbb{R}^d} |x - T(x)|^2 \rho^{k+1}(x)\, dx.$$

Then the optimal map T satisfies the Euler–Lagrange identity

$$\int_{\mathbb{R}^d} \left[\nabla \left(\frac{\delta \mathcal{F}}{\delta \rho}(\rho^{k+1}) \right)(x) + \frac{x - T(x)}{\tau} \right] \cdot \xi(x)\, \rho^{k+1}(x)\, dx = 0 \tag{26}$$

for every $\xi \in C_c^\infty(\mathbb{R}^d; \mathbb{R}^d)$. In particular,

$$T(x) = x + \tau \nabla \left(\frac{\delta \mathcal{F}}{\delta \rho}(\rho^{k+1}) \right)(x) \quad \text{for } \rho^{k+1}\text{-a.e. } x. \tag{27}$$

Proof. Let $\xi \in C_c^\infty(\mathbb{R}^d;\mathbb{R}^d)$ and define the perturbation map

$$\Phi_\varepsilon(x) := x + \varepsilon\xi(x), \qquad \rho_\varepsilon := (\Phi_\varepsilon)_\# \rho^{k+1}.$$

For $|\varepsilon|$ small, Φ_ε is a diffeomorphism (by smoothness and compact support of ξ).

Since ρ^{k+1} is a minimizer of the JKO functional, the function

$$J(\varepsilon) := \mathcal{F}(\rho_\varepsilon) + \frac{1}{2\tau} W_2^2(\rho_\varepsilon, \rho^k)$$

has a minimum at $\varepsilon = 0$. Hence

$$J'(0) = 0.$$

We compute the derivative of each term.

Step 1: Variation of $\mathcal{F}$. By Assumption (2),

$$\frac{d}{d\varepsilon}\Big|_{\varepsilon=0} \mathcal{F}(\rho_\varepsilon) = \int_{\mathbb{R}^d} \nabla\Big(\frac{\delta \mathcal{F}}{\delta \rho}(\rho^{k+1})(x)\Big) \cdot \xi(x)\, \rho^{k+1}(x)\, dx.$$

Step 2: Variation of the Wasserstein term. Let T be the optimal map from ρ^{k+1} to ρ^k (Assumption (3)). Define a map from ρ_ε to ρ^k by

$$T_\varepsilon := T \circ \Phi_\varepsilon^{-1}.$$

Indeed,

$$(T_\varepsilon)_\# \rho_\varepsilon = (T \circ \Phi_\varepsilon^{-1})_\# (\Phi_\varepsilon)_\# \rho^{k+1} = T_\# \rho^{k+1} = \rho^k.$$

Therefore, T_ε is an admissible transport map from ρ_ε to ρ^k, and thus

$$W_2^2(\rho_\varepsilon, \rho^k) \le \int_{\mathbb{R}^d} |\Phi_\varepsilon(x) - T(x)|^2\, \rho^{k+1}(x)\, dx.$$

At $\varepsilon = 0$, the right-hand side equals

$$\int |x - T(x)|^2 \rho^{k+1}(x)\, dx = W_2^2(\rho^{k+1}, \rho^k),$$

because T is optimal. Hence the one-sided derivative at $\varepsilon = 0$ satisfies

$$\frac{d}{d\varepsilon}\Big|_{\varepsilon=0} W_2^2(\rho_\varepsilon, \rho^k) = \frac{d}{d\varepsilon}\Big|_{\varepsilon=0} \int |\Phi_\varepsilon(x) - T(x)|^2 \rho^{k+1}(x)\, dx.$$

(One can justify equality by applying the same argument with $-\xi$ and using minimality at $\varepsilon = 0$.)

Now compute explicitly:

$$\Phi_\varepsilon(x) = x + \varepsilon\xi(x),$$

so

$$\frac{d}{d\varepsilon}\Big|_{\varepsilon=0} |\Phi_\varepsilon(x) - T(x)|^2 = 2(x - T(x)) \cdot \xi(x).$$

Therefore,

$$\frac{d}{d\varepsilon}\Big|_{\varepsilon=0} \frac{1}{2\tau} W_2^2(\rho_\varepsilon, \rho^k) = \frac{1}{\tau} \int_{\mathbb{R}^d} (x - T(x)) \cdot \xi(x)\, \rho^{k+1}(x)\, dx.$$

Step 3: First-order optimality. Since $J'(0) = 0$, combining Step 1 and Step 2 gives

$$\int_{\mathbb{R}^d} \left[\nabla\Big(\frac{\delta \mathcal{F}}{\delta \rho}(\rho^{k+1})\Big)(x) + \frac{x - T(x)}{\tau}\right] \cdot \xi(x)\, \rho^{k+1}(x)\, dx = 0$$

for every $\xi \in C_c^\infty(\mathbb{R}^d;\mathbb{R}^d)$, which is (26).

Since this holds for all test vector fields ξ, we conclude

$$\nabla\left(\frac{\delta \mathcal{F}}{\delta \rho}(\rho^{k+1})\right)(x) + \frac{x - T(x)}{\tau} = 0 \quad \text{for } \rho^{k+1}\text{-a.e. } x.$$

Rearranging yields (27). □

14 Closing summary

(1) Particle ODEs and density PDEs are microscopic and macroscopic descriptions of the same dynamics.

(2) Adding Brownian noise changes transport PDEs into Fokker–Planck equations.

(3) The score $\nabla \log \rho_t$ appears naturally when rewriting stochastic density evolution as a transport equation via probability current.

(4) Modern generative AI can be viewed as learning and integrating dynamics in probability space.

15 Flow Matching

Theorem 15.1 (Conditional velocity induces the correct marginal continuity equation)**.** *Let $(\Omega, \mathcal{G}, \mathbb{P})$ be a probability space. Let Z be an auxiliary random variable (e.g. $Z = (X_0, X_1)$), and let $(X_t)_{t\in[0,1]}$ be an $\mathbb{R}^d$-valued stochastic process such that:*

1. *For almost every sample ω, the path $t \mapsto X_t(\omega)$ is absolutely continuous.*
2. *There exists a measurable* conditional velocity *field $u_t(x \mid z) \in \mathbb{R}^d$ such that*

$$\frac{d}{dt}X_t = u_t(X_t \mid Z) \quad \textit{a.s. for a.e. } t \in [0,1]. \tag{28}$$

3. *For each t, X_t admits a density p_t on $\mathbb{R}^d$.*
4. *(Integrability) For every $\varphi \in C_c^\infty(\mathbb{R}^d)$,*

$$\mathbb{E}\big[\,|\nabla\varphi(X_t) \cdot u_t(X_t \mid Z)|\,\big] < \infty \quad \textit{for a.e. } t,$$

and differentiation under expectation is justified.

Define the marginal velocity field

$$v_t^\star(x) := \mathbb{E}\big[u_t(X_t \mid Z) \mid X_t = x\big]. \tag{29}$$

Then, for every test function $\varphi \in C_c^\infty(\mathbb{R}^d)$,

$$\frac{d}{dt}\int_{\mathbb{R}^d} \varphi(x)p_t(x)\,dx = \int_{\mathbb{R}^d} \nabla\varphi(x) \cdot v_t^\star(x)\,p_t(x)\,dx \quad \textit{for a.e. } t. \tag{30}$$

Equivalently, p_t solves the continuity equation

$$\partial_t p_t + \nabla \cdot (p_t v_t^\star) = 0 \tag{31}$$

in the weak sense.

Proof. Fix $\varphi \in C_c^\infty(\mathbb{R}^d)$. Since X_t has density p_t,

$$\mathbb{E}[\varphi(X_t)] = \int_{\mathbb{R}^d} \varphi(x) p_t(x)\, dx.$$

By the pathwise absolute continuity of X_t and the chain rule,

$$\frac{d}{dt}\varphi(X_t) = \nabla\varphi(X_t) \cdot \frac{d}{dt} X_t \quad \text{a.s. for a.e. } t.$$

Using (28),

$$\frac{d}{dt}\varphi(X_t) = \nabla\varphi(X_t) \cdot u_t(X_t \mid Z).$$

Taking expectation and using differentiation under the expectation (justified by Assumption 4),

$$\frac{d}{dt}\mathbb{E}[\varphi(X_t)] = \mathbb{E}\big[\nabla\varphi(X_t) \cdot u_t(X_t \mid Z)\big]. \tag{32}$$

Now apply the tower property (conditional expectation) with respect to X_t:

$$\mathbb{E}\big[\nabla\varphi(X_t) \cdot u_t(X_t \mid Z)\big] = \mathbb{E}\Big[\,\mathbb{E}\big[\nabla\varphi(X_t) \cdot u_t(X_t \mid Z) \mid X_t\big]\,\Big].$$

Since $\nabla\varphi(X_t)$ is measurable with respect to $\sigma(X_t)$, it can be pulled out of the inner conditional expectation:

$$= \mathbb{E}\Big[\nabla\varphi(X_t) \cdot \mathbb{E}[u_t(X_t \mid Z) \mid X_t]\Big].$$

By definition of $v_t^\star$ in (29),

$$= \mathbb{E}\big[\nabla\varphi(X_t) \cdot v_t^\star(X_t)\big].$$

Using the density p_t,

$$= \int_{\mathbb{R}^d} \nabla\varphi(x) \cdot v_t^\star(x)\, p_t(x)\, dx.$$

Combining this with (32) yields

$$\frac{d}{dt}\int_{\mathbb{R}^d} \varphi(x) p_t(x)\, dx = \int_{\mathbb{R}^d} \nabla\varphi(x) \cdot v_t^\star(x)\, p_t(x)\, dx,$$

which is (30).

Finally, integrating by parts (using compact support of φ),

$$\int_{\mathbb{R}^d} \nabla\varphi \cdot (p_t v_t^\star)\, dx = -\int_{\mathbb{R}^d} \varphi\, \nabla \cdot (p_t v_t^\star)\, dx.$$

Hence (30) is exactly the weak form of

$$\partial_t p_t + \nabla \cdot (p_t v_t^\star) = 0.$$

This proves (31). □

[Why the flow matching regression target is correct] Under the assumptions of Theorem 15.1, fix $t \in [0,1]$. Among all measurable vector fields $v : \mathbb{R}^d \to \mathbb{R}^d$, the minimizer of

$$\mathbb{E}\big[\|v(X_t) - u_t(X_t \mid Z)\|^2\big]$$

is given (uniquely up to p_t-a.e. equality) by

$$v(x) = v_t^\star(x) = \mathbb{E}[u_t(X_t \mid Z) \mid X_t = x].$$

Proof. This is the standard L^2 projection property of conditional expectation. For any $v(X_t)$ measurable with respect to $\sigma(X_t)$,

$$\mathbb{E}\big[\|u_t - v(X_t)\|^2\big] = \mathbb{E}\big[\|u_t - v_t^\star(X_t)\|^2\big] + \mathbb{E}\big[\|v_t^\star(X_t) - v(X_t)\|^2\big],$$

where the cross term vanishes by conditional expectation orthogonality. Hence the minimum is attained at $v = v_t^\star$. □