UNIT 7

A unified perspective on diffusion models

Markus Holzleitner, **Johannes Brandstetter**
Institute for Machine Learning

JOHANNES KEPLER
UNIVERSITY LINZ

Institute for
Machine Learning

Copyright statement:

Recap: Stochastic analysis

- Brownian motion: collection of RVs $w(t)$, $t \geq 0$, so that:
  - $w(0) = 0$
  - $w(t) \sim (0, t)$
  - Consider $\Delta w(t_1, t_2) = w(t_2) - w(t_1)$. Then for $t_1 < t_2 < t_3$: $\Delta w(t_1, t_2)$ is independent from $\Delta w(t_2, t_3)$
- Second-order BM-increments matter: $dw = \Delta w(t, t + dt)$, then $(dw)^2 = dt$
- Ito integral:

$$\int_0^t f(s)dw(s) \approx \sum_{i=0}^{n-1} f(s_i)\Delta w(s_i, s_{i+1})$$

- SDE on $[0, T]$:

$$dx(t) = f(x(t), t)dt + g(t)dw(t)$$

- Ito's lemma: chain rule to compute $d\phi(x_t, t)$ for given ϕ.

From SDEs to PDEs: Fokker-Planck (1)

- Aim: derive PDE characterizing time evolution of density $p_t(x)$ associated with the diffusion process defined by SDE

$$dx(t) = f(x(t), t)dt + g(t)dw(t)$$

  Focus on $D = 1$.

- Let $\phi(x, t)$ be a smooth test function . By Itô's lemma:

$$d\phi(x_t, t) = \left(\frac{\partial \phi}{\partial t} + \frac{\partial}{\partial x}\phi \cdot f(x_t, t) + \frac{1}{2} g^2(t) \frac{\partial^2}{\partial x^2}\phi \right) dt + g(t) \frac{\partial}{\partial x}\phi dw_t.$$

- Taking expectation over $p_t(x)$ and noting $\mathbb{E}[dw_t] = 0$ (Randomness in x_t comes from past noise):

$$\mathbb{E}[d\phi(x_t, t)] = \mathbb{E}\left[\left(\frac{\partial \phi}{\partial t} + \frac{\partial}{\partial x}\phi f(x_t, t) + \frac{1}{2} g^2(t) \frac{\partial^2}{\partial x^2}\phi \right) dt \right].$$

$$\mathbb{E}[d\phi(x_t, t)] = \int \left(\frac{\partial \phi}{\partial t} + \frac{\partial}{\partial x}\phi f(x, t) + \frac{1}{2} g^2(t) \frac{\partial^2 \phi}{\partial x^2} \right) p_t(x) dx dt.$$

From SDEs to PDEs: Fokker-Planck (2)

- Use integration by parts:

$$\int \frac{\partial \phi}{\partial x} f p_t dx = -\int \phi \frac{\partial}{\partial x}(f p_t) dx,$$
$$\int \frac{\partial^2 \phi}{\partial x^2} p_t dx = -\int \phi \frac{\partial^2 p_t}{\partial x^2} dx.$$

- Substituting back:

$$\mathbb{E}[d\phi(x_t, t)] = \int \phi(x, t) \left[\frac{\partial p_t}{\partial t} + \frac{\partial}{\partial x}(f p_t) - \frac{1}{2} g^2(t) \frac{\partial^2 p_t}{\partial x^2} \right] dx dt.$$

- Since ϕ is arbitrary, the integrand must vanish:

$$\int_0^T d\phi(x_t, t) = \phi(x_T, T) - \phi(x_0, 0) = 0$$

$$\frac{\partial p_t}{\partial t} = -\frac{\partial}{\partial x} \cdot (f(x, t) p_t(x)) + \frac{1}{2} g^2(t) \frac{\partial^2}{\partial x^2} p_t(x),$$

- If $g = 0$ (i.e. ODE-case): **transport equation**

From variational to score based (1)

- Assume forward transition

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

$$\epsilon = \mathcal{N}(0, 1)$$

- Define $\lambda_t = \log \frac{\alpha_t}{\sigma_t}$ for $t \in (0, T]$.
- Then the linear SDE

$$dx(t) = f(t)x(t)dt + g(t)dw(t),$$

with coefficients

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

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

has the given transitions.

- Converse also true.

From variational to score based (2)

- Time-dependent statistics:
  - Conditional mean $m(t) = \mathbb{E}(x_t | x(0) = x_0)$
  - Conditional covariance $P(t) = \mathrm{Cov}(x_t | x(0) = x_0)$.
- Associated evolution ODEs:
  - $\frac{dm(t)}{dt} = f(t)m(t)$
  - $\frac{dP(t)}{dt} = 2f(t)P(t) + g^2(t)\mathbf{I}$
- How? explicit solution for **linear SDE** of the form

$$dx(t) = f(t)x(t)dt + g(t)dw(t)$$

- Can be derived via **Ito's formula**.
- See e.g. The principles of diffusion models, Section C.1.5.

From variational to score based (3)

- Reverse process:

$$d\bar{x}(t) = \left[f(\bar{x}(t), t) - g^2(t) \underbrace{\nabla_x \log p_t(\bar{x}(t))}_{\text{Score}} \right] dt + g(t) d\bar{w}(t), \tag{1}$$

$$\bar{x}(T) \sim p_T, \quad \bar{w}(t) = w(T - t) - w(T)$$

- Forward diffusion spreads data into increasingly noisy configurations.
- How can reversing this process produce clean, structured samples concentrated near data manifold?
- **Diffusion term $g(t)d\bar{w}(t)$ coupled with score–driven drift $-g^2(t)\nabla_{\bar{x}} \log p_t(\bar{x}(t))$:**
  - Score guides trajectories toward regions of higher density
  - Noise allows controlled exploration.

From variational to score based (4)

- When $f = 0$, (1) reads:

$$\bar{x}(t) = -g^2(t)\nabla_x \log p_t\big(\bar{x}(t)\big)dt + g(t)d\bar{w}(t).$$

- Reparameterize time forward via $s = T - t$,rename BM: $d\bar{w}(t) = -dw_s$, write $\bar{x}_s = \bar{x}(T - s)$ and $\pi(s) = p_{T-s}$:

$$d\bar{x}(s) = g^2(T - s)\nabla \log \pi_s\big(\bar{x}_s\big)ds + g(T - s)dw(s)$$

- Tweedie's formula:

$$\mathbb{E}\left[x|\tilde{x}\right] = \tilde{x} + \sigma^2\nabla_{\tilde{x}} \log \pi_s(\tilde{x}).$$

  - Early on ($s \approx 0$, i.e., $t \approx T$), $g(T - s)$ is typically larger: process explores broadly.
  - As s increases: $g(T - s)$ decreases, score term dominates, pulling samples into high-density regions of π_s
  - $s = T$ (i.e., $t = 0$): trajectories close to data manifold.

- Rigorous derivation of (1) via Fokker-Planck: this blogpost

Recap: three main objectives

- **Variational View:** Learn a parametrized density $p_\theta(x_{t-\Delta t}|x_t)$ to approximate reverse transition $p(x_{t-\Delta t}|x_t)$ by minimizing:

$$\mathbb{E}_{p_t(x_t)}\left[\mathcal{D}_{\mathrm{KL}}\big(p(x_{t-\Delta t}|x_t)\|p_\theta(x_{t-\Delta t}|x_t)\big)\right]$$

- **Score-Based View:** Learn a score model $s_\phi(x_t, t)$ to approximate the marginal score $\nabla_x \log p_t(x_t)$ via:

$$\mathbb{E}_{p_t(x_t)}\left[\|s_\theta(x_t,t) - \nabla_x \log p_t(x_t)\|_2^2\right]$$

- **Flow-Based View:** Learn a velocity model $u_\theta(x_t, t)$ to match the oracle velocity $u_t(x_t)$ by minimizing:

$$\mathbb{E}_{p_t(x_t)}\left[\left\|u_\theta^t(x_t,t) - u^t(x_t)\right\|_2^2\right].$$

How to do in practice? Conditioning

- Based on the KL divergence in DDPMs:
  - ϵ-Prediction (Noise Prediction)

$$\mathbb{E}_t \left[\omega(t) \mathbb{E}_{x_0, \epsilon} \left\| \epsilon_\theta(x_t, t) - \epsilon \right\|_2^2 \right].$$

  - x-Prediction (Clean Prediction)

$$\mathbb{E}_t \left[\omega(t) \mathbb{E}_{x_0, \epsilon} \left\| x_\theta(x_t, t) - x_0 \right\|_2^2 \right].$$

- **Score Prediction**

$$\mathbb{E}_t \left[\omega(t) \mathbb{E}_{x_0, \epsilon} \left\| \mathbf{s}_\theta(x_t, t) - \nabla_{x_t} \log p_t(x_t | x_0) \right\|_2^2 \right],$$

- u-Prediction (Velocity Prediction)

$$\mathbb{E}_t \left[\omega(t) \mathbb{E}_{x_0, \epsilon} \left\| u_\theta(x_t, t) - u_t(x_t | x_0, \epsilon) \right\|_2^2 \right],$$

  with $u_t(x_t | x_0, \epsilon) = \alpha_t' x_0 + \sigma_t' \epsilon$.

- Conditional objectives differ by original ones only by shift!

Common denominators

- $$\mathbb{E}_{x_0,\epsilon} \underbrace{\mathbb{E}_{p_{\text{time}}(t)}}_{(A)} \Big[\underbrace{\omega(t)}_{(A)} ||\mathrm{NN}_\theta(\underbrace{x_t}_{(B)}, t) - \underbrace{(A_t x_0 + B_t \epsilon)}_{(C)} ||_2^2 \Big].$$
- Ad (B): all affine flows of the form $x_t = \alpha_t x_0 + \beta_t \epsilon$ mathematically equivalent. Can be transformed into each other by time-reparametrization and spatial rescaling.
- Ad (C): All regression targets (noise, clean, score based, flow based) of this form.
  - Can be transformed into each other easily.
  - Gradients of MSE objective are equivalent: differ only by possible time-reweighting factor.
- Ad (A): May improve training
- Details: The principles of diffusion models, Section 6

ODE flow and SDE produce same marginals

- Let $\gamma(t) \geq 0$ arbitrary. Let u be the vector field describing probability flow.
- Consider the reverse-time SDE

$$d\bar{x}(t) = \Big[u(\bar{x}(t), t) - \tfrac{1}{2}\gamma^2(t)s(\bar{x}(t), t)\Big]d\bar{t} + \gamma(t)d\bar{w}(t),$$

  evolving backward from $\bar{x}(T) \sim p_T$ down to $t = 0$.
- Process $\{\bar{x}(t)\}_{t\in[0,T]}$ matches prescribed marginals $\{p_t\}_{t\in[0,T]}$ induced by the ODE's density path.
- s and u can be linearly transformed into each other.
- Once marginal density path is fixed: an entire family of dynamics can reproduce it: Flow-ODE and reverse-time SDE above.
- Proof ($D = 1$): observe $\frac{\partial}{\partial x}(sp) = \frac{\partial^2}{\partial x^2}p$ to show $\partial_{\bar{t}}p = -\frac{\partial}{\partial x}(up)$

Conclusion

- Goal: common framework behind variational, score-based and flow-based perspective: deeply interconnected.
- Common source allowing stable and efficient learning: **conditioning**
- Evolution of probability densities: governed by same PDE: **Fokker-Planck**
- Various paremetrizations (noise, clean data, score, velocity) interchangeable.
- Prediction target matter of implementation and stability rather than fundamental modelling difference.
- **Ultimate takeaway:** all approaches learn time-dependent vector-field to transport simple prior to data distribution.