Machine Learning

Assignment 10

CHANG Yung-Hsuan (張永璿)

111652004

eiken.sc11@nycu.edu.tw

November 5, 2025

1. Consider a forward stochastic differential equation

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

show that the corresponding probability flow ordinary differential equation is written as

$$\mathrm{d}x_t = \left(f(x_t, t) - \frac{1}{2}\frac{\partial}{\partial x}(g(x,t))^2 - \frac{(g(x_t,t))^2}{2}\frac{\partial}{\partial x}\log p(x_t, t) \right)\mathrm{d}t.$$

Proof. Let the forward stochastic differential equation be

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

with density $p(x,t)$ satisfying the Fokker–Planck equation

$$\frac{\partial p}{\partial t} = -\frac{\partial}{\partial x}(fp) + \frac{1}{2}\frac{\partial^2}{\partial x^2}(g^2 p).$$

A deterministic ordinary differential equation $\mathrm{d}x_t = v(x_t, t)\,\mathrm{d}t$ transports p via the continuity equation

$$\frac{\partial p}{\partial t} = -\frac{\partial}{\partial x}(vp).$$

Equate the two right–hand sides and match fluxes, we have

$$\begin{aligned} -\frac{\partial}{\partial x}(vp) &= -\frac{\partial}{\partial x}(fp) + \frac{1}{2}\frac{\partial^2}{\partial x^2}(g^2 p) \\ &= -\frac{\partial}{\partial x}\left(fp - \frac{1}{2}\frac{\partial}{\partial x}(g^2 p) \right). \end{aligned}$$

Then,

$$vp = fp - \frac{1}{2}\frac{\partial}{\partial x}(g^2 p)$$

$$= fp - \frac{1}{2}\left(\left(\frac{\partial g^2}{\partial x}\right)p + g^2\frac{\partial p}{\partial x}\right).$$

Dividing both sides by p, we have

$$v(x,t) = f(x,t) - \frac{1}{2}\frac{\partial}{\partial x}(g(x,t))^2 - \frac{(g(x,t))^2}{2}\frac{\partial}{\partial x}\ln p(x,t),$$

where $\frac{\partial}{\partial x}\ln p = \frac{\frac{\partial p}{\partial x}}{p}$.

Hence the probability–flow ordinary differential equation is

$$\mathrm{d}x_t = \left(f(x_t,t) - \frac{1}{2}\frac{\partial}{\partial x}(g(x,t))^2 - \frac{(g(x_t,t))^2}{2}\frac{\partial}{\partial x}\ln p(x_t,t)\right)\mathrm{d}t.$$

2. **The Future of AI and the Foundations of Machine Learning**. Describe the future abilities of AI, machine learning models involved, and the modelization of the future abilities you proposed.

Answer. I pondered this problem while watching aircraft at the Taoyuan Airport. Some landed planes waited for free gates on the taxiway; this is just a waste of time. A perfect schedule could collapse due to weather, late arrivals, gate outages, or crew limits. The contrast was striking: the "plan" was deterministic, but the "system" was stochastic, governed by complex constraints like gate size, tow limits, and airline fairness. This made me wonder: *can we build learning-based algorithms that adapt to disruptions while ensuring feasibility and fairness?*

a. A Future AI Capability (20 years)

In 20 years, AI systems will design algorithms with formal guarantees for large, uncertain resource allocation problems like airport gates, seaport scheduling, power dispatch, or medical rostering. Today, learned heuristics adapt well but lack guarantees, while traditional optimization ensures guarantees but fails under uncertainty or scale.

b. Machine Learning Paradigms Involved

The backbone would be **supervised learning**, trained on historical and simulated high-quality solutions.

A lightweight **reinforcement-learning** layer would interact with real-time environments. When new tasks arrive (a flight delay, a power surge), it refines decisions through delay-aware rewards while respecting safety rules inherited from supervised models. **Self-supervised pretraining** over graph and sequence structures would further enable fast adaptation to unseen layouts, such as new terminals or shift structures.

c. Modelization

I cast the airport setting into a compact bin-packing model that stays testable yet faithful to operations. A "bin" is a gate/stand–time block with normalized capacity 1; an "item" is a flight turnaround whose size $s_i \in (0, 1]$ is its resource demand (e.g., minutes or crew load) divided by a base unit. We decide assignments $x_{i,b} \in \{0, 1\}$ with feasibility $\sum_i s_i x_{i,b} \leq 1$. The slack of bin b is $r_b = 1 - \sum_i s_i x_{i,b}$.

The objective is lexicographic: first minimize the number of used bins $\sum_b y_b$ (gate blocks actually opened); then, among all such solutions, minimize fragmentation via the squared-slack sum $\sum_b r_b^2$. This mirrors practice: operators primarily care about using as few gate blocks as possible (fewer remote stands, fewer tows), and secondarily about packing tightly to avoid tiny leftover gaps that waste resources.

This model is testable and comparable. For any instance, report `#bins`, the volume lower bound $\mathrm{LB}_{\mathrm{vol}} = \left\lceil \sum_i s_i \right\rceil$, and, on small sizes, the gap to the proven optimal value from branch-and-bound. Stress-test generalization with synthetic mixes (uniform, bimodal, near-0.5 "conflict", log-normal) and with size shifts (train at smaller n, evaluate at larger n), which emulate peak/off-peak patterns and irregular arrival waves.

As baselines, use best-fit decreasing and a bin-completion + subset-sum routine; these are fast and strong. For learning, use supervised **learn-to-rank** to score items from sizes and simple instance histograms, then `argsort` and pack with a deterministic best-fit decreasing backend; optionally, a small **learn-to-place** classifier selects among Top-K residual bins with a safe fallback to best-fit decreasing. A cost-aware greedy tie-break $\Delta_{\text{place}} = \lambda\left((r-s)^2 - r^2\right)$ vs. $\Delta_{\text{new}} = M + \lambda(1-s)^2$ with $M \gg 1$ keeps decoding aligned with the objective: good actions make Δ smaller (often negative), so squared-slack is minimized while bin count remains the primary goal.

3. There are unanswered questions from the lecture, and there are likely more questions we haven't covered. Take a moment to think about these questions. Write down the ones you find important, confusing, or interesting.

   Answer. We mentioned morphing in the lecture. I wonder how morphing with structures works; for example, mapping nose to nose, eyes to eyes, and mouth to mouth. In this case, does the probability-flow ordinary differential equation still apply?