ML Week10 Assignment

WANG TZU YI

November 11, 2025

1. Derivation of the Reverse SDE and Probability Flow ODE

Consider a general Itô SDE of the form

$$dX_t = f(X_t, t)\,dt + g(t)\,dW_t, \tag{1}$$

where $f(X_t, t)$ is the drift term, $g(t)$ is the diffusion coefficient (assumed independent of X_t), and W_t is a standard Wiener process.

1.1 Forward Fokker–Planck Equation

Let $p_t(x)$ denote the probability density function of X_t. The time evolution of $p_t(x)$ is governed by the Fokker–Planck equation:

$$\frac{\partial p_t(x)}{\partial t} = -\nabla_x \cdot \big(f(x,t)\,p_t(x)\big) + \frac{1}{2}\nabla_x \cdot \big(D(t)\,\nabla_x p_t(x)\big), \tag{2}$$

where $D(t) = g(t)g(t)^\top$ is the diffusion matrix. The probability current (or probability flux) is then defined as

$$J_t(x) = f(x,t)\,p_t(x) - \frac{1}{2}D(t)\,\nabla_x p_t(x), \tag{3}$$

so that the Fokker–Planck equation can be written compactly as

$$\frac{\partial p_t(x)}{\partial t} = -\nabla_x \cdot J_t(x). \tag{4}$$

1.2 Reverse-Time SDE

Now we consider running the diffusion process backward in time. Let $s = T - t$ and define $q_s(x) = p_{T-s}(x)$. Then

$$\frac{\partial q_s(x)}{\partial s} = -\frac{\partial p_t(x)}{\partial t} = \nabla_x \cdot J_t(x). \tag{5}$$

We now seek a reverse-time SDE of the form

$$dX_t = \tilde{f}(X_t, t)\, dt + \tilde{g}(t)\, d\bar{W}_t, \tag{6}$$

whose marginal density evolution matches the above. By comparing the two Fokker–Planck equations, and assuming $\tilde{g}(t) = g(t)$ (same diffusion strength), we obtain the relationship between the forward and reverse drifts:

$$\tilde{f}(x, t) = f(x, t) - D(t)\, \nabla_x \log p_t(x). \tag{7}$$

Thus, the reverse-time SDE becomes

$$\boxed{dX_t = \left[f(X_t, t) - g^2(t)\, \nabla_x \log p_t(X_t)\right] dt + g(t)\, d\bar{W}_t.} \tag{8}$$

1.3 Associated Probability Flow ODE

We can now define a deterministic ODE that yields the same time-dependent marginal distributions $p_t(x)$. Let the ODE be

$$dX_t = v(X_t, t)\, dt, \tag{9}$$

whose corresponding continuity equation is

$$\frac{\partial p_t}{\partial t} = -\nabla_x \cdot (v(x, t)\, p_t(x)). \tag{10}$$

To match the Fokker–Planck equation (2), we require that

$$v(x, t)\, p_t(x) = f(x, t)\, p_t(x) - \frac{1}{2} D(t)\, \nabla_x p_t(x),$$

which implies

$$v(x, t) = f(x, t) - \frac{1}{2} D(t)\, \nabla_x \log p_t(x). \tag{11}$$

Hence, the deterministic **probability flow ODE** is

$$\boxed{dX_t = \left[f(X_t, t) - \tfrac{1}{2} g^2(t)\, \nabla_x \log p_t(X_t)\right] dt.} \tag{12}$$

1. The Future Capability of AI: Towards Genuine Thinking

While today's AI systems excel at pattern recognition, reasoning imitation, and linguistic generation, they fundamentally lack the ability to *think*. Current large language models (LLMs) predict words based on statistical patterns rather than forming causal reasoning chains or internal self-awareness.

Twenty years from now, I believe AI will evolve toward what I call a **Causal Cognitive Reasoning AI**—a system capable of performing human-like thought processes that are interpretable, self-corrective, and causally grounded. Such an AI would not simply answer questions but would:

- Formulate hypotheses autonomously (as in scientific reasoning);
- Derive and test causal relationships between observations;
- Detect contradictions and revise its own reasoning strategies;
- Explain its thinking process in natural, interpretable language.

Application scenarios. In science, such an AI could autonomously generate new research hypotheses, reconcile conflicting theories, and simulate experiments to refine its internal models of the world. In education, students could learn reasoning and metacognition by engaging in dialogue with an AI that transparently reveals how it thinks. In philosophy and ethics, such systems may deepen our understanding of rationality, responsibility, and consciousness itself.

Ultimately, this shift from "language imitation" to "thought co-creation" would mark a fundamental transformation in the role of AI—elevating it from a tool to an intellectual partner for humanity.

2. The Machine Learning Paradigms Involved

The realization of thinking AI will require an integration of multiple learning paradigms, primarily:

- **Supervised learning** for fundamental linguistic and logical structure understanding;
- **Unsupervised learning** for discovering latent concepts and semantic networks from vast text corpora;
- **Reinforcement learning** for exploring reasoning pathways and refining strategies through internal feedback.

Reasoning. The data source for such systems would consist of both external text data (human knowledge) and internal reasoning traces (the model's own generated "thoughts"). The *target signal* would not be a correct answer, but the coherence, causal validity, and explainability of its reasoning chains. The model must receive feedback through *self-reflective reinforcement*, where logical consistency or contradiction reduction functions as an internal reward.

This hybrid paradigm allows the AI not only to map input-output relations, but to continually evaluate, challenge, and refine its internal representations—an essential step toward genuine cognition.

3. First Modeling Step: Simulating Human Thought Flow

As an initial research step, I propose constructing a simplified model problem called the **Thought-Trace Simulation Task**. Instead of generating direct answers, a large language model would be trained to produce explicit reasoning chains that mirror human thought flow:

> *Q: Why do leaves turn yellow?*
> *AI: (Step 1) Photosynthesis decreases → (Step 2) Chlorophyll breaks down → (Conclusion) Leaf color fades as chlorophyll declines.*

The AI would then internally evaluate the causal and logical soundness of each step, refining its reasoning through feedback.

Testability. Model performance can be quantitatively measured via a *Reasoning Consistency Score*, which evaluates:

1. Logical self-consistency of the reasoning chain;
2. Factual correctness relative to external knowledge bases;
3. Structural similarity to human-labeled reasoning steps.

Success would be defined by the AI's ability to sustain coherent reasoning under multiple perspectives while maintaining interpretability.

Required mathematical and ML tools.

- Extended Transformer architectures with memory and meta-cognitive feedback loops;
- **Causal inference** for explicit modeling of cause-effect relationships;
- **Reinforcement learning** for reasoning-path optimization;
- **Contrastive learning** for distinguishing valid from invalid reasoning trajectories.

This simplified problem embodies the long-term goal of constructing AI systems that not only generate text but simulate structured, interpretable human thought processes.

4. Conclusion

If in twenty years AI can truly *think*—constructing causal reasoning, self-reflecting, and explaining its thoughts—it will cease to be a mere computational instrument. It will become an **intellectual collaborator**, capable of co-discovering knowledge and extending human cognition itself. Such an achievement would redefine not only artificial intelligence but also the very notion of thinking.

References

- Lake, B. M., Ullman, T., Tenenbaum, J. B., and Gershman, S. (2017). *Building machines that learn and think like people.* Behavioral and Brain Sciences.
- Bengio, Y. (2023). *From System 1 Deep Learning to System 2 Deep Learning.* NeurIPS Keynote.
- Song, Y., and Ermon, S. (2021). *Score-Based Generative Modeling through SDEs.*