Preferred Networks

Mar. 9 2026 Amica Conference:
Thermodynamic Approaches to Artificial Intelligence.

Learning as a Finite-Time Non-Equilibrium Process

Daisuke Okanohara
Preferred Networks Inc. / Matlantis Corp.

Why This Question Now?

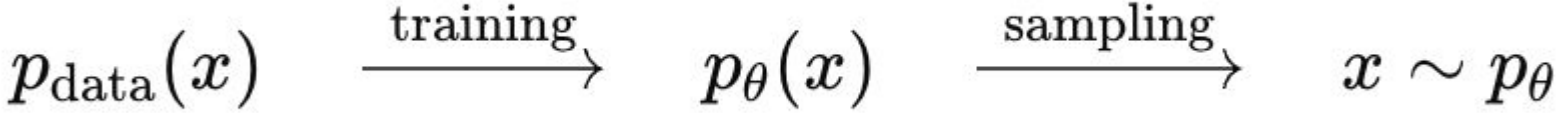
- Empirical progress in AI is extremely rapid.
  - Large language models, Image and video generation etc.
- But theoretical understanding lags behind.
  - e.g., why continual learning is difficult

What is learning in physical terms?

This presentation is based on
D. Okanohara “A Thermodynamic Theory of Learning I: Irreversible Ensemble Transport and Epistemic Costs”, arXiv:2601.17607
D. Okanohara “A Thermodynamic Theory of Learning Part II: Critical Period Closure and Continual Learning Failure”, arXiv:2602.07950

Modern Generative AI

Generative AI = Learning and sampling
high-dimensional probability distributions



- Extremely high-dimensional state space (text, images, molecules)
- Parametric model with 10^9 ~ 10^{12} parameters
- Objective: $p_\theta(x) \approx p_{\text{data}}(x)$

Diffusion Models as Finite-Time Non-Equilibrium Transport

Forward (Entropy ↑)

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

$$\partial_t p_t = -\nabla \cdot (f p_t) + \frac{1}{2} g^2 \Delta p_t$$

Relaxation toward simple equilibrium

Reverse (Learned Transport)

$$dx = \left[f(x,t) - g(t)^2 \nabla_x \log p_t(x) \right] dt + g(t)\, d\bar{W}_t$$

Score-based timereversal

Diffusion models implement finite-time nonequilibrium transport in **data space**

But learning itself is also a dynamical process.
What distribution evolves during learning ?

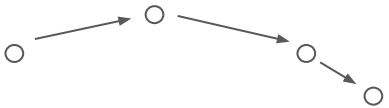
4

Learning as Ensemble Transport in Parameter Space

Standard View of Learning

- Single initialization and trajectory
- Point dynamics in parameter space

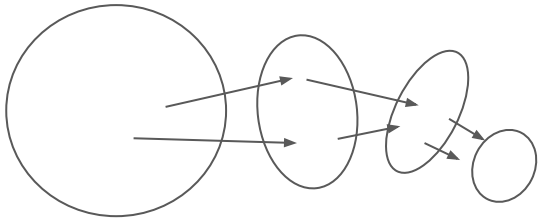
$$\theta_0 \to \theta_t$$



Ensemble View (for analysis)

- Random initialization and stochasticity generate an ensemble
- Distributional dynamics in parameter space

$$p_0(\theta) \to p_t(\theta)$$



Learning defines a dynamical system on **parameter space.**
We should study how distribution evolve.

Ensemble View of Learning

Parameter Distribution $q_s(\theta)$

distribution over parameters at learning time s.

Learning as Transport $\partial_s q_s + \nabla \cdot (q_s v_s) = 0$

Continuity equation in parameter space.

Point dynamics $d\theta_s = -\nabla\Phi(\theta_s)\, ds + \sqrt{2T}\, dW_s$

Φ: Objective (Langevin)

Distributional dynamics $\partial_s q_s = \nabla \cdot (q_s \nabla\Phi(\theta)) + T\, \Delta q_s$

(Fokker-Planck)

Epistemic Free Energy

$$\mathcal{F}[q] := \underbrace{\mathbb{E}_q[\Phi]}_{\text{ensemble-averaged objective}} - \underbrace{TH[q]}_{\text{entropy}}$$

- Objective minimization under entropy regularization.
- Φ: Objective function
- T: effective noise temperature

Note: This is **not ELBO**. It is defined over parameter distribution
This is not Bayes NN. This is defined as a bookkeeping quantity

Learning Irreversibility and Dissipation

Free Energy Dissipation

$$\frac{d}{ds}F[q_s] = -\sigma_s$$

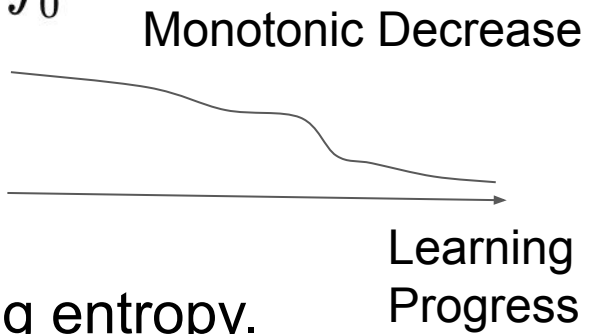
$$\sigma_s = \int q_s(\theta)\,\|v_s(\theta)\|^2 d\theta \geq 0$$

- Quadratic transport cost
- Measures irreversible probability flow

Total Entropy Production

$$F[q_0] - F[q_1] = \Sigma_{0:1}$$

$$\Sigma_{0:1} = \int_0^1 \sigma_s ds$$



Learning reduces free energy by producing entropy.

Finite-time learning is intrinsically **irreversible**.

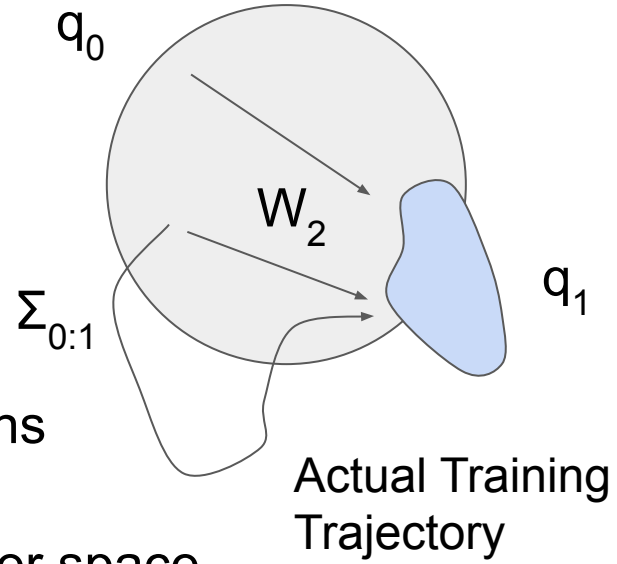
Preferred Networks

Epistemic Speed Limit

Epistemic Speed Limit

$$\mathcal{F}[q_0] - \mathcal{F}[q_1] = \Sigma_{0:1}$$

$$\Sigma_{0:1} \geq W_2(q_0, q_1)^2$$



W_2: Wasserstein-2 distance

$q_0 q_1$: initial and final parameter distributions

Measures geometric displacement in parameter space

Entropy production lower-bounds transport distance

All finite-time learning obeys a geometric speed limit

Consequence of the Epistemic Speed Limit

From $F[q_0] - F[q_1] = \Sigma_{0:1}$ and $\Sigma_{0:1} \geq W_2(q_0, q_1)^2$

We obtain:

$= \Sigma_{0:1} + T(H[q_0] - H[q_1])$

$$\underbrace{\mathbb{E}_{q_0}[\Phi] - \mathbb{E}_{q_1}[\Phi]}_{\text{Objective improvement}} \geq \underbrace{W_2(q_0, q_1)^2}_{\text{Transport}} + \underbrace{T\big(H[q_0] - H[q_1]\big)}_{\text{Hypothesis commitment}}$$

- Objective improvement requires geometric displacement
- Entropy reduction adds additional cost
- Learning cannot improve “for free”

Improvement is constrained by geometry and dissipation

Reachable Set Under Finite-Time Learning

Finite-Time Constraint

$$W_2(q_0, q_1) \leq \sqrt{F[q_0] - F[q_1]}$$

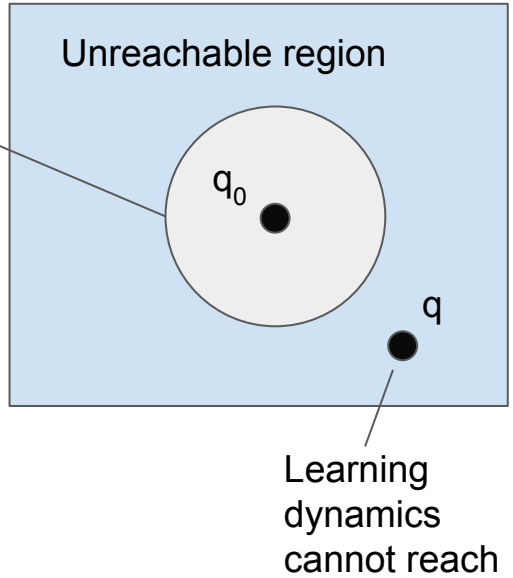
- Free energy reduction is finite
- Therefore, displacement is bounded

Reachable Set

$$q_1 \in \mathcal{B}_{W_2}\left(q_0, \sqrt{F[q_0] - F[q_1]}\right)$$

Reachable ensembles lie inside a Wasserstein ball.

Wasserstein ball in distribution space (point=dist.)

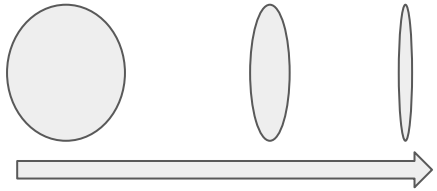


Preferred Networks

Difficulty in Continual Learning

Why Continual Learning Failure is Inevitable

Finite-time learning produces dissipation.
→ Excess dissipation induces anisotropic contraction.



Distribution contract

Ψ□ : transport map in parameter space

$$\theta_t = \Psi_t(\theta_0)$$

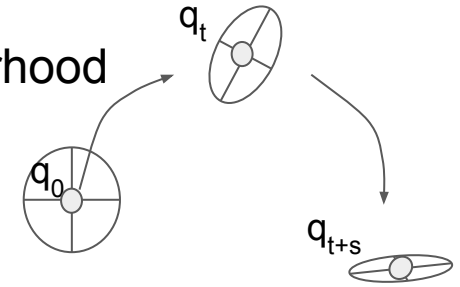
J□ : local deformation of directions

$$J_t = \partial \Psi_t / \partial \theta_0$$

Composition

$$\Psi_{t+s} = \Psi_s \circ \Psi_t \quad J_{t+s} = J_s J_t$$

q_0 and
the neighborhood



Submultiplicativity → irreversible collapse of reachable directions

Once collapsed early,
it cannot be recovered later.

Continual Learning as a Geometric Compatibility Problem

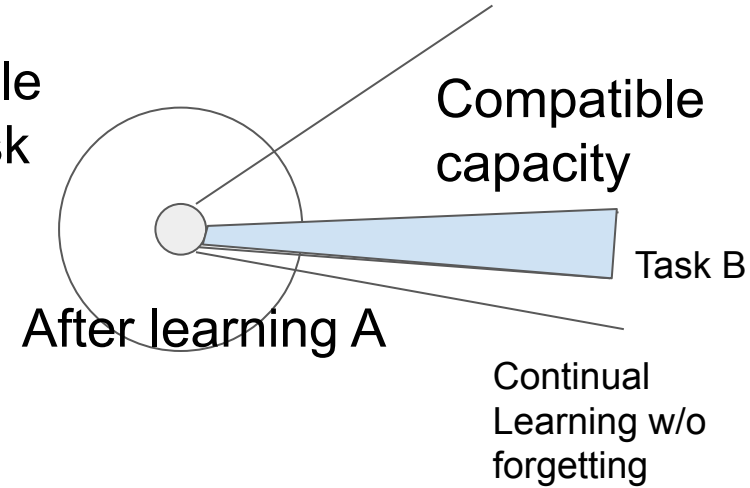
Continual Learning Task A → Task B

After learning A:

- Compatible capacity: $_A(t)$
  - task-preserving directions remain usable
  - preserve task A while adapting new task

New Task B requires effective dimension m_B

If $m_B > R_A(t)$ → forgetting is inevitable



Preferred Networks

Summary

Summary:
Learning as Finite-Time Non-Equilibrium Transport

Learning as Finite-Time Transport

- Learning evolves a parameter distribution q_s
- Free energy decreases under dissipation

$$\mathcal{F}[q] := \mathbb{E}_q[\Phi] - TH[q] \qquad \Sigma_{0:1} \geq W_2(q_0, q_1)^2$$

Finite-time improvement requires geometric displacement.

Learning composes multiplicatively

$$J_{t+s} = J_s J_t$$

Compatible directions cannot increase

Future work

Structured parameter space / Functional Space

- In this first step, we intentionally use a simplified setting, though in practice representation learning and inter-layer dynamics are crucial.

Relation to Symmetry

- Invariance, equivalence are important for keeping internal “compatible degrees of freedom”