All you need is Relative Information

Shaowei Lin

20230626

Singular Learning Theory
and Alignment Summit

Journey

- **2008**. Dream - "Never separate Memory from Compute."
- **2009**. Singular learning (with Bernd Sturmfels, Mathias Drton, Sumio Watanabe)

- **2011**. SLT - "All you need is *relative* information."
- **2012**. Spiking networks (with Chris Hillar)

- **2016**. AlphaGo - "Inference without alignment is broke or brute."
- **2017**. Dependent type theory and program synthesis

- **2020**. DTT - "Information/energy is constructive."
- **2021**. Category theory and information cohomology (with Chris Hillar)

Projects

1. **Information Cohomology**. (today; with Chris Hillar, Tai-Danae Bradley)

2. **Spiking Networks**. (today; with Chris Hillar, Sarah Marzen)

3. **Program Synthesis**. (inference with alignment)

   - Domain-specific languages for LLMs and RLs
   - Categorical proof assistants and tactics
   - Generalized algebraic theories and type-classes

Topos Institute is hiring!
(`shaowei@topos.institute`)

State Density

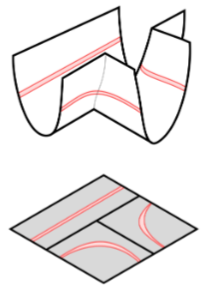
Level sets of relative information $H(s)$ unlock everything else.

Density of states

$$\nu(E) = \int \delta(H(s) - E)\,\mathrm{d}s$$

Two-sided Laplace transform

Mellin transform



$$Z = \int \nu(E)\, e^{-E}\, dE = \int e^{-H(s)}\,\mathrm{d}s$$

Partition function

$$\zeta(z) = \int H(s)^z\,\mathrm{d}s$$

Zeta function

[1] Jesse Hoogland, "Physics I: The Thermodynamics of Learning",Singular Learning Theory and Alignment Summit 2023.

Part I

Spiking neural networks

Statistical Learning

Setup.

- Observed variable X
- Hidden variable Z
- True distribution $q(X)$
- Model distribution $p_\theta(X, Z)$ parametrized by θ
- Marginal distribution $p_\theta(X) = \int p_\theta(X, Z) dZ$

Goal.

- Find θ minimizing

$$I_{q\|p_\theta}(X) = \int q(X) \log \frac{q(X)}{p_\theta(X)} dX$$

Variational Inference

Trick.

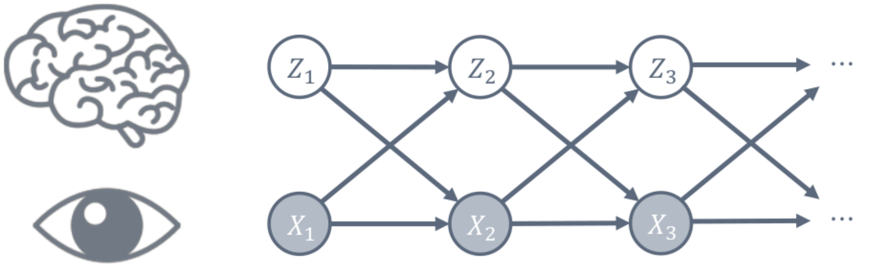
- Introduce distribution $q(Z|X)$ as extra parameter for optimization
- *Discriminative* distribution $q(X, Z) = q(Z|X)q(X)$
- *Generative* distribution $p_\theta(X, Z) = p_\theta(X|Z)p_\theta(Z)$ (usually)
- Minimize
$$I_{q||p_\theta}(X, Z) = \int q(X, Z) \log \frac{q(X, Z)}{p_\theta(X, Z)} dXdZ$$
by alternatingly varying q while holding p_θ fixed and vice versa.

Variants.

- *EM algorithm* (Dempster-Laird-Rubin)[2]. Let $q(Z|X)$ be $p_\theta(Z|X)$ at each step of the optimization.
- *em algorithm* (Amari)[3]. Let $q(Z|X)$ be parametrized $q_\lambda(Z|X)$ and alternatingly optimize θ and λ.
- Amari's *em* algorithm is biologically more plausible because **Bayesian inversion is hard**!

[2] Dempster, A.P., N.M. Laird, and D.B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm." JRSS 39, no. 1 (1977): 1-22.
[3] Amari, Shun-ichi. "Information geometry of the EM and em algorithms for neural networks." Neural networks 8, no. 9 (1995): 1379-1408.

Time Series with Memory

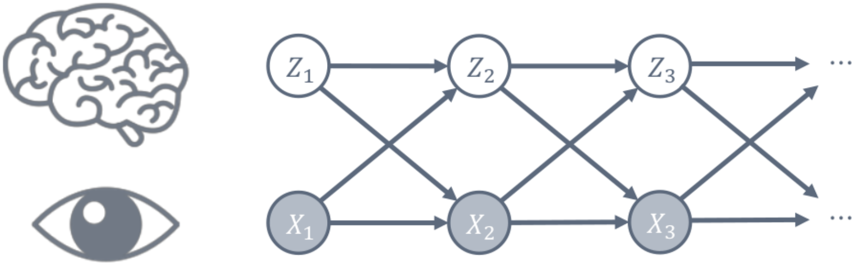


- **Time.** Assume discrete time for simplicity.
- **Environment**. $X_1, X_2, \ldots$. Immutable. Possibly partially hidden.
- **Memory**. $Z_1, Z_2, \ldots$. Mutable. Not latent/hidden variables!

- **Goal**. Optimize use of limited memory for predicting environment.
- **Objective.** Minimize

$$\lim_{T \to \infty} \frac{1}{T} I_{q\|p_\theta}(X_{1\ldots T}, Z_{1\ldots T})$$

Information Constraints



Put different constraints on structure of $q(Z_{1\ldots T}|X_{1\ldots T}) = \prod_k q(Z_{k+1}|Z_{1\ldots k}, X_{1\ldots T})$.
Let p^* denote the resulting p_θ that minimizes $I_{q\|p_\theta}(X_{1\ldots T}, Z_{1\ldots T})$.

- **No constraints.** $q(Z_{1\ldots T}|X_{1\ldots T}) = \prod_k q(Z_{k+1}|Z_{1\ldots k}, X_{1\ldots T})$. Optimal $I_{\text{free}} = I_{q\|p^*}(X_{1\ldots T})$.
- **Online learning.** $q(Z_{1\ldots T}|X_{1\ldots T}) = \prod_k q(Z_{k+1}|Z_{1\ldots k}, X_{1\ldots k})$. Optimal $I_{\text{online}} > I_{\text{free}}$.
- **Limited memory.** $q(Z_{1\ldots T}|X_{1\ldots T}) = \prod_k q(Z_{k+1}|Z_k, X_k)$. Optimal $I_{\text{mem}} > I_{\text{online}}$.

Relative Information Rate

- Assume limited memory, i.e. Markov process $q(Z_{1\dots T}|X_{1\dots T}) = \prod_k q(Z_{k+1}|Z_k, X_k)$.
- Assume q has stationary distribution $\bar{\pi}$. Let $\bar{q}$ be same Markov process but with initial $\bar{\pi}$.
- Using Kingman's subadditive ergodic theory[4] and under mild regularity conditions[5],

$$\lim_{T\to\infty} \frac{1}{T} I_{q\|p}(X_{1\dots T}, Z_{1\dots T}) = I_{\bar{q}\|p}(Z_2, X_2|Z_1, X_1).$$

- In continuous-time, we get the *relative information rate*

$$\lim_{T\to\infty} \frac{1}{T} I_{q\|p}(X_{1\dots T}, Z_{1\dots T}) = \frac{d}{dt} I_{\bar{q}\|p}(X_{1\dots 1+t}, Z_{1\dots 1+t})\Big|_{t=0}.$$

[4] https://en.wikipedia.org/wiki/Kingman%27s_subadditive_ergodic_theorem

[5] Brian G Leroux. "Maximum-likelihood estimation for hidden markov models." Stochastic processes and their applications, 40(1):127–143, 1992.

Stochastic Approximation

Setup. Parametric models $q_\lambda(Z_{n+1}|Z_n, X_n)$ and $p_\theta(Z_{n+1}, X_{n+1}|Z_n, X_n)$.

Goal. Minimize conditional relative information $I_{\bar{q}||p}(Z_2, X_2|Z_1, X_1)$.

Stochastic Approximation.[6]

1. Sample environment X_{n+1} from true distribution $q(X_{n+1}|X_n)$.
2. Sample memory Z_{n+1} from discriminatory distribution $q_\lambda(Z_{n+1}|Z_n, X_n)$.
3. Sample the generator gradient $\nabla_\theta I_{\bar{q}||p}(Z_2, X_2|Z_1, X_1)$ using Z_{n+1}, X_{n+1}.
4. Sample the discriminator gradient $\nabla_\lambda I_{\bar{q}||p}(Z_2, X_2|Z_1, X_1)$ using Z_{n+1}, X_{n+1}.
5. Update parameters θ, λ and repeat until convergence.

[6] Robbins, Herbert, and Sutton Monro. "A stochastic approximation method." The annals of mathematical statistics (1951): 400-407.

Stochastic Gradients

Generator.

$$\begin{aligned}&\nabla_\theta I_{\bar{q}\|p}(Z_2, X_2 | Z_1, X_1)\\ &= \lim_{T\to\infty} \mathbb{E}_q \left[\nabla_\theta \log p_\theta(Z_{T+1}, X_{T+1} | Z_T, X_T)\right]\end{aligned}$$

Discriminator.

$$\begin{aligned}&\nabla_\lambda I_{\bar{q}\|p}(Z_2, X_2 | Z_1, X_1)\\ &= \lim_{T\to\infty} \mathbb{E}_q \left[\underbrace{\left(\sum_{i=1}^{T} \nabla_\lambda \log q_\lambda(Z_{i+1} | Z_i, X_i) \right)}_{\text{momentum}} \underbrace{\log \frac{q_\lambda(Z_{T+1}, X_{T+1} | Z_T, X_T)}{p_\theta(Z_{T+1}, X_{T+1} | Z_T, X_T)}}_{\text{surprise}} \right]\end{aligned}$$

- Use discounted momentum (scale summands by some β^{T-i} with $\beta<1$) for numerical stability.
- Same as reinforcement learning with surprise as reward (policy gradient for average reward)[7].

[7] Karimi, Belhal, Blazej Miasojedow, Eric Moulines, and Hoi-To Wai. "Non-asymptotic analysis of biased stochastic approximation scheme." PMLR 2019 pp. 1944-1974.

Spiking Neural Networks

- Work over *continuous-time* instead of discrete-time, but concepts are the same.
- Optimize *relative information rate* instead of conditional relative information.
- Many neuron models possible[8].
    - Neuron spikes are described as Poisson processes controlled by cell potentials.
    - Potentials increase with incoming spikes, and reset with outgoing spikes.
    - Cell potentials and synaptic credit assignments decay with time.
- Stochastic approximation explains the triplet rule in spike-time-dependent plasticity!
- Discriminator surprise seems to explain dopamine-based neuromodulation!
- Discounted momentum seems to explain neuronal adaptation and refractoriness!

[8] https://shaoweilin.github.io/posts/2021-06-05-spiking-neural-networks/

Part II

Information Cohomology

Big Picture

- For simplicity, consider measures (need not be probabilistic) over finite sets.
- Think of the comparison between a model measure $\mathbb{P}$ and a true measure $\mathbb{Q}$ as a functor $F : P \to Q$ between categories of *weighted contexts and substitutions*.
- We want to a measure of how good F is at modeling truth at each morphism f in P. Define *relative information* as a functor $I_F : P \to \mathcal{R}$, where $\mathcal{R}$ is the category of *dual numbers*.
- Given $f : X \to Y$ in P, let P_f be the subcategory containing just f. The functors $P_f \to \mathcal{R}$ which are localized at Q, form an algebra A. Derivations (1-cocycles) $\delta : A \to A$ are generated by relative information!

Relative Information for Measures

- Let p and q be two measures on a finite set X with the same total measure

$$\sum_{x \in X} p(x) = \sum_{x \in X} q(x).$$

- Let $\pi : Y \to X$ be a measure-preserving map,
  i.e. there exists $p(y|x)$ for each $y \in Y$ and $x = \pi(y)$ such that $p(y) = p(y|x)p(x)$ and

$$\sum_{y \in \pi^{-1}(x)} p(y|x) = 1 \quad \text{for all } x,$$

  and similarly for q.

- Define the (conditional) relative information to be

$$I_{p \rightsquigarrow q}(\pi) = \sum_{x \in X} q(x) \sum_{y \in \pi^{-1}(x)} q(y|x) \log \frac{q(y|x)}{p(y|x)}.$$

Contexts and Substitutions

- A *context* is a finite set.

- A *substitution* $f : X \to Y$ between contexts is a set map $\pi_f : Y \to X$, together with conditional probabilities $p_f(y|x) \geq 0$ on Y for each $x \in X$, such that $p_f(y|x) = 0$ if $\pi(y) \neq x$ and $\sum_{y \in Y} p_f(y|x) = 1$.

- Two substitutions $f : X \to Y$ and $g : Y \to Z$ *compose* with set maps and conditional probabilities

$$\pi_{g \circ f} = \pi_f \circ \pi_g$$
$$p_{g \circ f}(z|x) = p_g(z|y)\, p_f(y|x).$$

- For each context X, there is an *identity* substitution $\mathrm{id} : X \to X$ with the identity set map π_{id} and the conditional probability $p_{\mathrm{id}}(x|x) = 1$.

- A *trivial context* $*$ is a one-element set. Substitutions $* \to X$ give probabilities on X.

Weighted Contexts and Substitutions

- A *weighted* context is a context X with a measure $p_X(x)$ on X.
- A *weighted* substitution $f : X \to Y$ is a substitution that is measure-preserving[9], i.e.
$$p_Y(y) = p_f(y|x)p_X(x) \quad \text{for all } x, y.$$
- Addition $X \oplus Y$ of weighted contexts is the disjoint union of underlaying sets and measures. Addition $f \oplus g$ of weighted substitutions is the disjoint union of underlying maps and conditionals. Check that the disjoint union of conditionals is again a conditional.
- Multiplication $X \otimes Y$ of weighted contexts is the product of underlying sets and measures. Multiplication $f \otimes g$ of weighted substitutions is the product of the underlying maps and conditionals. Check that the product of conditionals is again a conditional.

[9] Baez, John C., and Tobias Fritz. "A Bayesian characterization of relative entropy." arXiv preprint arXiv:1402.3067 (2014).

Dual Numbers

- The rig (semiring) of *duals* is $\mathcal{R} = \mathbb{R}_{\geq 0}[\varepsilon]/\langle \varepsilon^2 \rangle$, where ε is an infinitesimal with $\varepsilon^2 = 0$. Denote addition by $\oplus$ and multiplication by $\otimes$.
- We may also use the *extended duals* $\mathcal{R}_\infty = \mathbb{R}_{\geq 0,\infty}[\varepsilon]/\langle \varepsilon^2 \rangle$, where $\mathbb{R}_{\geq 0,\infty}$ has $\infty + a = \infty$ for all a; $\infty \times a = \infty$ for all $a \neq 0$; and $\infty \times 0 = 0$.
- We also think of the duals (extended duals) as a category, where the objects are reals (extended reals) a, and the morphisms are also reals (extended reals) $b : a \to a$ that compose by addition.
- Check that addition $\oplus$ and multiplication $\otimes$ extends to the objects and morphisms. In particular, if $b : a \to a$ and $d : c \to c$, then

$$b \otimes d : (a \times c) \to (a \times c)$$
$$b \otimes d = a \times d + b \times c$$

Information Categories and Relative Information

- An *information category* is a *rig* category (with $\oplus, \otimes$) of weighted contexts and substitutions. Think of an information category as a joint distribution on a collection of random variables.
- Given information categories P ("model distribution") and Q ("true distribution"), we compare them using a functor $F : P \to Q$.
- For each context X in P, we define the *total measure*

$$I_F(X) = \sum_{x \in X} p(x) = \sum_{x \in X} q(x)$$

- For each morphism $f : X \to Y$ in P, we define the *conditional* relative information

$$I_F(f) = \sum_{x \in X} q(x) \sum_{y \in \pi^{-1}(x)} q(y|x) \log \frac{q(y|x)}{p(y|x)}$$

- We call I_F the *relative information*.

Relative Information as a Functor

- Relative information is a functor $I_F : P \to \mathcal{R}$ from the information category P to the dual numbers $\mathcal{R}$ localized at Q.

- Total measure: let X and Y be objects in P.

  - **Measure preservation**. Morphisms in $\mathcal{R}$ are self-loops $a \to a$,
    so we must have $I_F(X) = I_F(Y)$ for all morphisms $f : X \to Y$ in P.

  - **Sum rule**. Measures of disjoint unions are sums of measures.

$$I_F(X \oplus Y) = I_F(X) \oplus I_F(Y)$$

  - **Product rule**. Measures of products are products of measures.

$$I_F(X \otimes Y) = I_F(X) \otimes I_F(Y)$$

Relative Information as a Functor

- Relative information is a functor $I_F : P \to \mathcal{R}$ from the information category P to the dual numbers $\mathcal{R}$ localized at Q.
- Conditional relative information: let $f : X \to Y, g : Y \to Z, h : X' \to Y'$ be morphisms in P.
  - **Chain rule**. Information of compositions are sums of information.

$$I_F(g \circ f) = I_F(g) + I_F(f).$$

  - **Sum rule**. Information of disjoint unions are sums of information.

$$I_F(f \oplus h) = I_F(f) \oplus I_F(h) = I_F(f) + I_F(h)$$

  - **Product rule**. Information of products behave like *derivations*. [10]

$$I_F(f \otimes h) = I_F(f) \otimes I_F(h) = I_F(X) \times I_F(h) + I_F(f) \times I_F(X')$$

  - **Localization**. If $p(y|x) = q(y|x)$ for all x, y, then $I_F(f) = 0$.

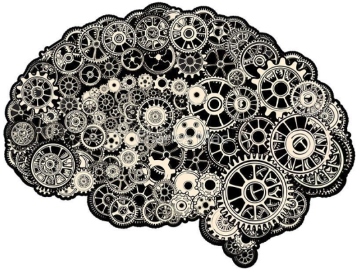
[10] Bradley, Tai-Danae. "Entropy as a topological operad derivation." Entropy 23, no. 9 (2021): 1195.

Relative Information as Cohomology

- For each $f : X \to Y$ in P, let P_f be the subcategory with objects X, Y and morphism f.
- The functors $P_f \to \mathcal{R}$ which are localized at Q, form an algebra A.
- Linear maps $\delta : A \to A$ satisfy all the conditions on the last two slides, except for the product rule of conditional relative information. Relative information I_F is one such linear map.
- **Conjecture**. Relative information I_F is a cocycle in the Hochschild cohomology of A.
- **Questions**.
  1. What are the higher-order cocycles? [11]
  2. Generalize to other kinds of information categories?
  3. State densities? Partition functions? Zeta functions?

[11] Baudot, Pierre, and Daniel Bennequin. "The homological nature of entropy." Entropy 17, no. 5 (2015): 3253-3318.

Thank you!



shaoweilin.github.io