

# RELATIVE INFORMATION AND THE DUAL NUMBERS

**Shaowei Lin**  
**Topos Institute**  
**(with Chris Hillar)**

20250110

Joint Mathematics Meetings: AMS Special Session on  
Algebraic Methods in Machine Learning and Optimization

## RELATIVE INFORMATION

- ▶ Given probability distributions  $q$  and  $p$  on a *finite* set  $\mathcal{E}_X$ , the *relative information* (Kullback-Leibler divergence, relative entropy) from  $p$  to  $q$  is

$$I_{q\|p} := \sum_{x \in \mathcal{E}_X} q(x) \log \frac{q(x)}{p(x)}.$$

- ▶ Given probability densities  $q$  and  $p$  on an *infinite* set  $\mathcal{E}_X$ , the relative information is

$$I_{q\|p} := \int q(x) \log \frac{q(x)}{p(x)} dx.$$

- ▶ Well-defined only when  $p(x) = 0$  implies  $q(x) = 0$  for all  $x$  (*absolute continuity*  $q \ll p$ ).
- ▶ Think of  $q$  as the *reference* or *true* distribution, and we want to know the distance of a *model* distribution  $p$  to the truth. This distance is not symmetric, i.e.  $I_{q\|p} \neq I_{p\|q}$ .

## INFORMATION IS RELATIVE!

- ▶ Relative information  $I_{q\|p}$  is well-defined for large classes of statistical models. Entropy  $H_p$ , on the other hand, is often ill-defined. In fact, when defined, we have

$$H_p = I_{\Delta_p\|pp} = \iint \Delta_p(x, y) \log \frac{\Delta_p(x, y)}{p(x)p(y)} dx dy = \int p(x) \log \frac{1}{p(x)} dx$$

where  $\Delta_p(x, y) = \mathbb{I}_{x=y} p(x)$  and  $pp(x, y) = p(x)p(y)$  are distributions on  $\mathcal{E}_X \times \mathcal{E}_X$ .

- ▶ To remind myself that information in a distribution should always be measured relative to another, I use the mantra: **INFORMATION IS RELATIVE!**
- ▶ Generally, let  $q, p$  be finite measures on a measurable space  $(\mathcal{E}_X, \mathcal{B}_X)$  with *total measure*  $T_q = T_p$ . If  $q \ll p$ , let  $dq/dp$  be the Radon-Nikodym derivative. Define the *relative information*

$$I_{q\|p} := \int dq \log \frac{dq}{dp} = T_q I_{\bar{q}\|\bar{p}}$$

where  $\bar{q}, \bar{p}$  are the normalized measures with  $T_{\bar{q}} = T_{\bar{p}} = 1$ .

## MOTIVATION: SINGULAR LEARNING

Let  $\{p(\cdot|\omega), \omega \in \Omega\}$  be a parametric model (a family of distributions) on  $X$ .

Let  $\varphi(\omega)$  be a prior on the parameter space  $\Omega$ . Let  $q$  be the true distribution of  $X$ .

Suppose we observe data  $x_{[n]} = (x_1, \dots, x_n) \in X^n$ .

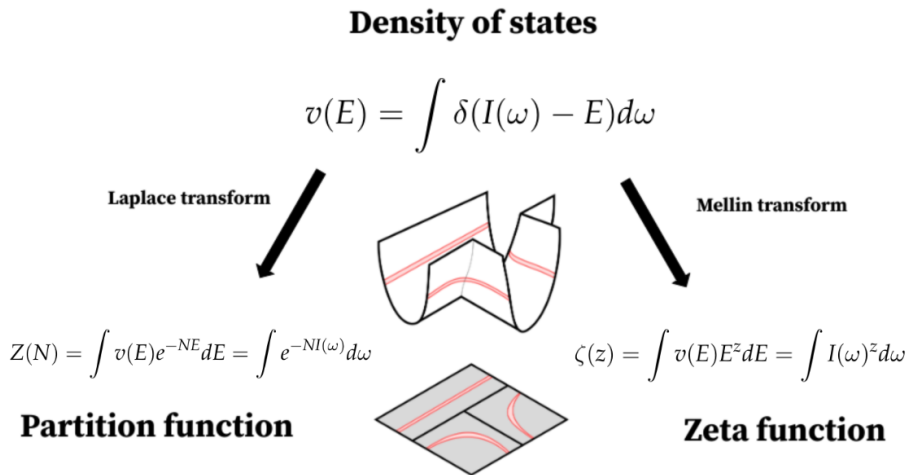
### Theorem (Convergence of stochastic complexity - Watanabe)

The *stochastic complexity* has an asymptotic expansion (as  $n \rightarrow \infty$ )

$$-\log Z_n = nS_n + \lambda_q \log n - (\theta_q - 1) \log \log n + O_p(1)$$

where  $\lambda_q$  is the *real log canonical threshold* of relative information  $I(\omega)$  over  $\Omega$  with respect to  $\varphi(\omega)$ , and  $\theta_q$  its multiplicity. For regular models, this is the Bayesian Information Criterion.

# MOTIVATION: SINGULAR LEARNING



<sup>1</sup>Jesse Hoogland, "Physics I: The Thermodynamics of Learning", Singular Learning Theory and Alignment Summit 2023.

## MOTIVATION

What properties of relative information uniquely define it?

How do we generalize relative information to non-statistical settings?

## CONDITIONAL RELATIVE INFORMATION (PROBABILITIES)

- ▶ Consider joint probabilities  $q(y, x)$  for  $(y, x) \in Y \times X$ .
- ▶ Conditional probabilities are  $q(y|x) := q(y, x)/q(x)$  when  $q(x) := \sum_y q(y, x) \neq 0$ .
- ▶ Given distributions  $q, p$  on  $Y \times X$ , the *conditional* relative information from  $p$  to  $q$  is

$$I_{q\|p}(Y|X) := \sum_{x \in X} q(x) \sum_{y \in Y} q(y|x) \log \frac{q(y|x)}{p(y|x)}.$$

- ▶ Important concept for variational inference, expectation-maximization algorithm.

## CONDITIONAL RELATIVE INFORMATION (MEASURES)

- ▶ More generally, let  $\mathcal{M}_Z = (\mathcal{E}_Z, \mathcal{B}_Z)$  and  $\mathcal{M}_X = (\mathcal{E}_X, \mathcal{B}_X)$  be measurable spaces. Assume that  $\mathcal{B}_Z$  contains the singletons  $\{z\}$  for all  $z \in \mathcal{E}_Z$ , and similarly for  $\mathcal{B}_X$ . Let  $\pi : \mathcal{M}_Z \rightarrow \mathcal{M}_X$  be a *measurable surjection*. Think of  $\mathcal{M}_Z$  as a *refinement* of  $\mathcal{M}_X$ .
- ▶ Given a finite measure  $q$  on  $\mathcal{M}_Z$ , the pushforward  $q_\pi(dx) := d(\pi_*q)$  gives a measure on  $\mathcal{M}_X$ . The pullback  $\pi^*\mathcal{B}_X$  is a sub  $\sigma$ -algebra of  $\mathcal{B}_Z$ , so the *conditional expectation*  $\mathbb{E}_q[\cdot | \pi^*\mathcal{B}_X] : \mathcal{E}_Z \rightarrow \mathbb{R}$  exists for any measurable  $f : \mathcal{E}_Z \rightarrow \mathbb{R}$ . The preimages  $\pi^{-1}(x), x \in \mathcal{E}_X$ , generate  $\mathcal{B}_Z$ , so  $\mathbb{E}_q[f | \pi^*\mathcal{B}_X]$  is constant on each preimage. Given  $x \in \mathcal{E}_X$ , define the *conditional measure*  $q_\pi(dz|x)$  on  $\mathcal{M}_Z$  by

$$q_\pi(B|x) := \mathbb{E}_q[\mathbb{I}_B | \pi^*\mathcal{B}_X](z) \quad \text{for all } B \in \mathcal{B}_Z \text{ and any } z \in \pi^{-1}(x).$$

- ▶ Given finite measures  $q, p$  on  $\mathcal{M}_Z$  with  $T_q = T_p$  and  $q \ll p$ , let  $q_\pi(dz|x)/p_\pi(dz|x)$  denote the Radon-Nikodym derivative. The *conditional relative information* is

$$I_{q||p}(Z|X) := \int q_\pi(dx) \int q_\pi(dz|x) \log \frac{q_\pi(dz|x)}{p_\pi(dz|x)}.$$



## CHAIN RULE

Let  $\pi : \mathcal{M}_Z \rightarrow \mathcal{M}_X$  be a measurable surjection. Let  $q, p$  be finite measures on  $\mathcal{M}_Z$ . By abuse of notation, we also let  $q, p$  denote their pushforwards on  $\mathcal{M}_X$ .

**Theorem (Chain Rule for Total Measure)**

$$T_q(Z) = T_q(X)$$

**Theorem (Chain Rule for Relative Information)**  $I_{q\|p}(Z) = I_{q\|p}(Z|X) + I_{q\|p}(X)$

**Proof (for probability measures over finite sets)**

$$\begin{aligned} I_{q\|p}(Y, X) &= \sum_{x,y} q(y, x) \log \frac{q(y, x)}{p(y, x)} \\ &= \sum_{x,y} q(y|x)q(x) \log \frac{q(y|x)q(x)}{p(y|x)p(x)} \\ &= \sum_{x,y} q(y|x)q(x) \log \frac{q(y|x)}{p(y|x)} + \sum_{x,y} q(y|x)q(x) \log \frac{q(x)}{p(x)} = I_{q\|p}(Y|X) + I_{q\|p}(X) \end{aligned}$$

## SUM AND PRODUCT RULES (TOTAL MEASURE)

- ▶ Suppose we have measure spaces  $(\mathcal{E}_X, \mathcal{B}_X, \mu_X)$  and  $(\mathcal{E}_Y, \mathcal{B}_Y, \mu_Y)$  with finite  $\mu_X, \mu_Y$ .
- ▶ Let the sum  $\mathcal{E}_X + \mathcal{E}_Y$  be the disjoint union  $\mathcal{E}_X \sqcup \mathcal{E}_Y$ .  
Let the sum  $\mathcal{B}_X + \mathcal{B}_Y$  be the collection of  $B \subseteq \mathcal{E}_X + \mathcal{E}_Y$  such that  $B \cap \mathcal{E}_X \in \mathcal{B}_X, B \cap \mathcal{E}_Y \in \mathcal{B}_Y$ .  
Let the sum  $\mu_X + \mu_Y$  satisfy  $\mu_X + \mu_Y(B) = \mu_X(B \cap \mathcal{E}_X) + \mu_Y(B \cap \mathcal{E}_Y)$  for all  $B \in \mathcal{B}_X + \mathcal{B}_Y$ .
- ▶ Let the product  $\mathcal{E}_X \times \mathcal{E}_Y$  be the Cartesian product of sets.  
Let the product  $\mathcal{B}_X \times \mathcal{B}_Y$  be the  $\sigma$ -algebra generated by  $B_X \times B_Y$  for all  $B_X \in \mathcal{B}_X, B_Y \in \mathcal{B}_Y$ .  
Let the product  $\mu_X \times \mu_Y$  satisfy  $\mu_X \times \mu_Y(B_X \times B_Y) = \mu_X(B_X)\mu_Y(B_Y)$  for all  $B_X \in \mathcal{B}_X, B_Y \in \mathcal{B}_Y$ .
- ▶ Total measures satisfy the sum and product rules.

$$T_{\mu_X + \mu_Y} = T_{\mu_X} + T_{\mu_Y}$$

$$T_{\mu_X \times \mu_Y} = T_{\mu_X} T_{\mu_Y}$$

## SUM AND PRODUCTS (RELATIVE INFORMATION)

For relative information, we also have sum and product rules. For each  $i \in \{1, 2\}$ , let  $q_i$  and  $p_i$  be finite measures on  $(\mathcal{E}_{Y_i}, \mathcal{B}_{Y_i})$  and  $(\mathcal{E}_{X_i}, \mathcal{B}_{X_i})$  respectively, with  $T_{q_i} = T_{p_i}$ .

### Theorem (Sum Rule)

$$I_{(q_1+q_2)\|(p_1+p_2)}(Y_1 + Y_2|X_1 + X_2) = I_{q_1\|p_1}(Y_1|X_1) + I_{q_2\|p_2}(Y_2|X_2)$$

[ similar to  $d(f + g) = df + dg$  ]

### Theorem (Product Rule)

$$I_{(q_1 \times q_2)\|(p_1 \times p_2)}(Y_1 \times Y_2|X_1 \times X_2) = T_{q_2} I_{q_1\|p_1}(Y_1|X_1) + T_{q_1} I_{q_2\|p_2}(Y_2|X_2)$$

[ similar to  $d(fg) = g df + f dg$  ]

# AXIOMATIZATION OF RELATIVE INFORMATION

- ▶ We see that relative information satisfies the chain, sum and product rules.
- ▶ Under appropriate conditions (e.g. continuity), the only functions on probabilities that satisfy those rules<sup>2</sup> are scalar multiples of relative information. There are similar axiomatization results for classical and quantum entropy. See papers and talks below for more information.
  - Baez, Fritz, Leinster. "A characterization of entropy in terms of information loss." Entropy 13(11), 2011.
  - Baez, Fritz. "A Bayesian characterization of relative entropy." arXiv:1402.3067, 2014.
  - Baudot, Bennequin. "The homological nature of entropy." Entropy 17(5), 2015.
  - Vigneaux. "Information structures and their cohomology." arXiv:1709.07807, 2017.
  - Bradley. "Entropy as a topological operad derivation." Entropy 23(9), 2021.
  - Maszczyk. "Hochschild cohomology for abstract convexity and Shannon entropy." [youtu.be/Zt9xO56CBG0](https://youtu.be/Zt9xO56CBG0), 2023.

---

<sup>2</sup>Turns out that the chain and sum rules are enough. The product rule is a consequence of these two rules.

## RIGS AND MEASURE SPACES

- ▶ A *rig category*  $\mathbf{C}$  is one that has a symmetric monoidal structure  $(\mathbf{C}, +, 0)$  for *addition*, a monoidal structure  $(\mathbf{C}, \times, 1)$  for *multiplication*, and some natural isomorphisms for distributivity and annihilation, that together satisfy some coherence laws. A *rig functor*  $\mathcal{F} : \mathbf{C} \rightarrow \mathbf{D}$  is one that preserves the rig structure on  $\mathbf{C}$  and  $\mathbf{D}$ .
- ▶ Let **Mble** be the category whose objects are *measurable spaces*  $\mathcal{M}_X = (\mathcal{E}_X, \mathcal{B}_X)$  and whose morphisms  $\pi : \mathcal{M}_Z \rightarrow \mathcal{M}_X$  are measurable surjections  $\pi_{\mathcal{E}} : \mathcal{E}_Z \rightarrow \mathcal{E}_X$ .
- ▶ Let **Meas** be the category whose objects are *measure spaces*  $\mathcal{M}_X = (\mathcal{E}_X, \mathcal{B}_X, \mu_X)$  and whose morphisms  $\pi : \mathcal{M}_Z \rightarrow \mathcal{M}_X$  are measurable surjections satisfying  $\mu_X = (\pi_{\mathcal{E}})_* \mu_Y$ .
- ▶ Both **Mble** and **Meas** are rig categories under disjoint union  $+$  and Cartesian product  $\times$ .

## INFORMATION RIGS

- ▶ A *poset* is a category where between any two objects, there is at most one morphism.
- ▶ An *information rig* is a pair  $(\mathbf{P}, \mathcal{M})$  for some *rig poset*  $\mathbf{P}$  and some *rig functor*  $\mathcal{M} : \mathbf{P} \rightarrow \mathbf{Mble}$ .
- ▶ To introduce finite measures  $q, p$  for comparison in  $I_{q||p}$ , we need a clever way to assign measures to objects of  $\mathbf{P}$  in a consistent way that avoids explicit tracking of pushforwards.
- ▶ Let  $\mathcal{U} : \mathbf{Meas} \rightarrow \mathbf{Mble}$  be the *forgetful functor* that ignores the measures.
- ▶ Let  $(\mathbf{P}, \mathcal{M})$  be an information rig.
  - We say that a rig functor  $q : \mathbf{P} \rightarrow \mathbf{Meas}$  is a *lift* of  $\mathcal{M}$  if  $\mathcal{U} \circ q = \mathcal{M}$ .
  - We say that  $q$  is *finite* if  $q$  maps each  $X \in \text{Ob } \mathbf{P}$  to a measure space with finite measure  $q_X$ .
  - We say that  $q, p$  have the *same total measure* if  $T_{q_X} = T_{p_X}$  for all  $X \in \text{Ob } \mathbf{P}$ .
  - We say that  $q \ll p$  if  $q_X \ll p_X$  for all  $X \in \text{Ob } \mathbf{P}$ .

## THE RIG OF DUAL NUMBERS

- ▶ Let  $\bar{\mathbb{R}}_{\geq 0}$  be the extended nonnegative reals  $[0, \infty]$  as a rig/semiring with addition  $+$  and multiplication  $\times$ , satisfying  $a + \infty = \infty$  for all  $a$ ,  $a \times \infty = \infty$  for all  $a \neq 0$ , and  $0 \times \infty = 0$ .
- ▶ The rig of *duals* is  $\mathcal{R} = \bar{\mathbb{R}}_{\geq 0}[\varepsilon]/\langle \varepsilon^2 \rangle$ . Think of  $\varepsilon$  is an infinitesimal with  $\varepsilon^2 = 0$ .
- ▶ We shall think of the rig of duals as a *category*  $\mathbf{R}$ , where
  - the extended nonnegative reals  $a \in \bar{\mathbb{R}}_{\geq 0}$  are *objects*;
  - the duals  $a + b\varepsilon \in \mathcal{R}$  are *morphisms* from  $a$  to itself, i.e. loops;
  - the morphisms *compose* by tangent addition  $(a + b\varepsilon) \circ (a + c\varepsilon) = a + (b + c)\varepsilon$ ;
  - the dual  $a + 0\varepsilon \in \mathcal{R}$  is the *identity* morphism from  $a$  to itself.
- ▶ The category  $\mathbf{R}$  is a *rig category* under addition  $+$  and multiplication  $\times$ .
  - $(a + b\varepsilon) + (c + d\varepsilon)$  is the morphism  $(a + c) + (b + d)\varepsilon$  from the object  $a + c$  to itself.
  - $(a + b\varepsilon) \times (c + d\varepsilon)$  is the morphism  $(ac) + (ad + bc)\varepsilon$  from the object  $ac$  to itself.

## RELATIVE INFORMATION FROM A RIG FUNCTOR

- Fix an information rig  $(\mathbf{P}, \mathcal{M})$ . Assume that the finite rig functors  $q, p : \mathbf{P} \rightarrow \mathbf{Meas}$  are lifts of  $\mathcal{M}$ , have the same total measure and satisfy  $q \ll p$ .

- For each morphism  $\pi : Z \rightarrow X$  in  $\mathbf{P}$ , denote the total measure by

$$T_q(Z) := T_{q_Z} \quad T_q(X) := T_{q_X} \quad T_q(\pi) := T_{q_Z} = T_{q_X}.$$

- For each morphism  $\pi : Z \rightarrow X$  in  $\mathbf{P}$ , let  $q_\pi(dx) := q_X(dx)$  and  $q_\pi(dz|x)$  be the associated conditional measure. Denote the relative information by

$$I_{q\|p}(\pi) := \int q_\pi(dx) \int q_\pi(dz|x) \log \frac{q_\pi(dz|x)}{p_\pi(dz|x)}.$$

### Theorem

Let  $F_{q\|p} : \mathbf{P} \rightarrow \mathbf{R}$  map each object  $X$  to the real number  $T_q(\pi)$ , and each morphism  $\pi : Z \rightarrow X$  to the dual number  $T_q(\pi) + I_{q\|p}(\pi)\varepsilon$ . Then  $F_{q\|p}$  is a *rig functor*.



# RELATIVE INFORMATION FROM A RIG FUNCTOR

## Proof Outline

Claims about total measure.

- ▶ Check that  $F_{q||p}$  maps morphisms  $\pi : Z \rightarrow X$  in  $\mathbf{P}$  to loops  $a \rightarrow a$  in  $\mathbf{R}$ , i.e.

$$T_q(Y) = T_q(X) = a.$$

- ▶ Check that  $F_{q||p}$  maps disjoint unions of objects in  $\mathbf{P}$  to sums of reals in  $\mathbf{R}$ , i.e.

$$T_q(X_1 + X_2) = T_q(X_1) + T_q(X_2).$$

- ▶ Check that  $F_{q||p}$  maps Cartesian products of objects in  $\mathbf{P}$  to products of reals in  $\mathbf{R}$ , i.e.

$$T_q(X_1 \times X_2) = T_q(X_1) T_q(X_2).$$

Indeed, the claims follow the chain rule, sum rule and product rule for total measure.

# RELATIVE INFORMATION FROM A RIG FUNCTOR

## Proof Outline

Claims about relative information.

- Check that  $F_{q||p}$  maps compositions in  $\mathbf{P}$  to tangent sums in  $\mathbf{R}$ , i.e.

$$I_{q||p}(\pi_1 \circ \pi_2) = I_{q||p}(\pi_1) + I_{q||p}(\pi_2).$$

- Check that  $F_{q||p}$  maps disjoint unions of morphisms in  $\mathbf{P}$  to sums of duals in  $\mathbf{R}$ , i.e.

$$I_{q||p}(\pi_1 + \pi_2) = I_{q||p}(\pi_1) + I_{q||p}(\pi_2).$$

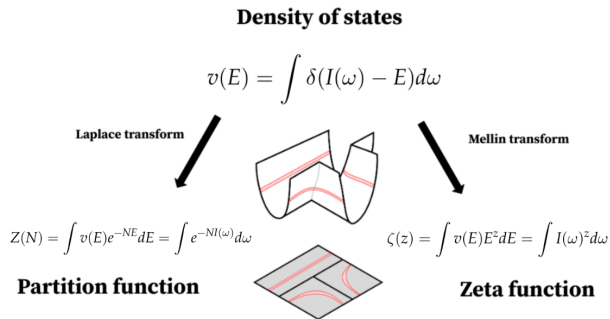
- Check that  $F_{q||p}$  maps Cartesian products in  $\mathbf{P}$  to products in  $\mathbf{R}$ , i.e.

$$I_{q||p}(\pi_1 \times \pi_2) = T_q(\pi_2) \cdot I_{q||p}(\pi_1) + T_q(\pi_1) \cdot I_{q||p}(\pi_2).$$

Indeed, the claims follow from the chain, sum and product rules for relative information.

# WHY RELATIVE INFORMATION?

- Generalized relative information from rig functors, from cohomology
- Beautiful algebra, geometry and combinatorics

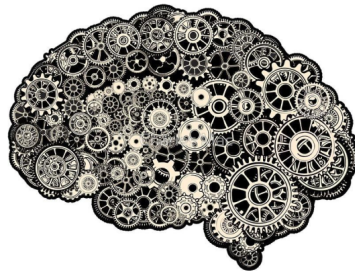


- Information is relative! Information is energy! It from bit! <sup>4</sup>

<sup>3</sup>Jesse Hoogland, "Physics I: The Thermodynamics of Learning", Singular Learning Theory and Alignment Summit 2023.

<sup>4</sup>Wheeler, J.A. (1989). Information, physics, quantum: the search for links. Int Symp on Foundations of Quantum Mechanics. Tokyo: pp. 354-358.

Thank you!



`shaoweilin.github.io`