


# The Complexity of User Retention

**Eli Ben-Sasson**

Department of Computer Science, Technion, Haifa, Israel


[eli@cs.technion.ac.il](mailto:eli@cs.technion.ac.il)

 <https://orcid.org/0000-0002-0708-0483>

**Eden Saig**

Department of Computer Science, Technion, Haifa, Israel

[edens@cs.technion.ac.il](mailto:edens@cs.technion.ac.il)

 <https://orcid.org/0000-0002-0810-2218>

---

## Abstract

This paper studies families of distributions  $\mathcal{T}$  that are amenable to *retentive learning*, meaning that an expert can *retain* users that seek to predict their future, assuming user attributes are sampled from  $\mathcal{T}$  and exposed gradually over time. Limited *attention span* is the main problem experts face in our model. We make two contributions.

First, we formally define the notions of *retentively learnable* distributions and properties. Along the way, we define a *retention complexity* measure of distributions and a natural class of *retentive scoring rules* that model the way users evaluate experts they interact with. These rules are shown to be tightly connected to truth-eliciting “proper scoring rules” studied in Decision Theory since the 1950’s [McCarthy, PNAS 1956].

Second, we take a first step towards relating retention complexity to other measures of significance in computational complexity. In particular, we show that linear properties (over the binary field) are retentively learnable, whereas random Low Density Parity Check (LDPC) codes have, with high probability, maximal retention complexity. Intriguingly, these results resemble known results from the field of property testing and suggest that deeper connections between retentive distributions and locally testable properties may exist.

**2012 ACM Subject Classification** Theory of computation → Models of computation

**Keywords and phrases** retentive learning, retention complexity, information elicitation, proper scoring rules

**Digital Object Identifier** 10.4230/LIPIcs.ITCS.2019.12

**Related Version** <https://eccc.weizmann.ac.il/report/2017/160/>

**Funding** This work was supported by the Israeli Science Foundation under grant number 1501/14.

**Acknowledgements** We thank Yuval Filmus for many helpful discussions. We thank the anonymous reviewers for their insightful comments on an earlier version of this paper.

## 1 Introduction

Certain aspects of life – child development, love, psychological disorders, etc. – seem comprehensible and somewhat predictable *only* when addressed jointly with an expert, via an interactive process that unfolds over time. This work initiates the formal study of phenomena that are amenable to an interactive *collaborative discovery* process between an expert and a layperson. This collaboration unfolds over time, as more aspects of the phenomena become apparent to the layperson. We model this by associating each layperson with a sequence



© Eli Ben-Sasson and Eden Saig;  
licensed under Creative Commons License CC-BY  
10th Innovations in Theoretical Computer Science (ITCS 2019).

Editor: Avrim Blum; Article No. 12; pp. 12:1–12:30



Leibniz International Proceedings in Informatics

LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

$x = (x_1, \dots, x_n)$  sampled from some underlying distribution  $\mathcal{T}$ , and assume that attributes of the phenomena (entries of  $x$ ) are gradually exposed over time, in response to active querying by the layperson. The interactive process has the expert predicting the value of  $x_i$  before it is revealed to the layperson. At the time of revelation, the layperson also re-evaluates his assessment of the expert’s utility to him: outcomes that seem “obvious” and predictable to the user are viewed as less helpful than “surprising” and unpredictable ones. At the end of each step, the layperson must decide whether to terminate his relationship with the expert or continue with it.

A major factor in our study is the *limited attention span* that characterizes users; consequently, the expert’s goal is to *retain* the layperson by constantly supplying him with meaningful and surprising insights about the phenomena. The terms *guru* (instead of expert) and *follower* (as opposed to layperson) better capture this dynamic so we use these terms henceforth. The main advantage the guru has over her followers is a greater ability to understand the phenomena, which we model by assuming the guru has a greater *memory span* than her followers. Using these terms, we define a class of distributions to be *retentively learnable* if a fixed *constant* advantage in memory span suffices for a guru to retain her followers throughout the entire collaborative discovery process. Our main complexity result is that the property of *linearity* over the binary field is retentively learnable (Theorem 10), whereas arbitrary linear spaces are maximally non-retentive (Theorem 13). To state these results, we first have to define our model and study its intrinsic properties. These properties turn out to be non-trivial to study, and lead to surprising results that are of independent interest. In particular, our first main result (Theorem 4) is a non-intuitive characterization of retentive scoring rules in terms of *proper scoring rules* studied in Decision Theory.

## Roadmap

Subsection 1.1 further explains, informally, the kind of guru–follower dynamics that we aim to formalize and understand. Subsection 1.2 gives a formal description of the model. In Subsection 1.3, we discuss the concept of *retentive scoring*, which allows us to describe the collaborative discovery process in an incentive-compatible way, taking agent rationality into account. Subsection 1.4 adds the layer of *limited memory span* to characterize the discrepancies between the guru and follower, and between fellow gurus. In Subsection 1.5, we state our main complexity results for linear properties. Subsection 1.6 reviews the latest related work, and finally Subsection 1.7 summarizes the main contributions and questions to be explored in the future.

## 1.1 The Guru’s Problem

In today’s information society, crowd-based automated gurus gather data from users on a voluntary basis in order to produce meaningful insights. The quality of insights greatly depends on the amount and quality of data provided by the users, but those users have limited attention, giving rise to the study of *attention economy* [12, 16]. By asking “*interesting questions*” and making “*meaningful predictions*”, an automated guru can retain users, but only if it “knows” how to ask “interesting” questions and provide “meaningful” feedback.

The phenomenon that motivated this research is that of *early child development*; the gurus are experts in this field and the followers are parents of newborn babies [3]. For the sake of concreteness we shall continue using this particular setting to describe our model but it may be conveniently replaced by the reader with physicians or psychologists playing the gurus as they interact with patients (followers) regarding a complex medical or mental

problem, or with financial advisors as gurus and their follower clientele. In these and similar settings, gurus and followers discuss a complex phenomenon that evolves over time, which the followers wish to understand, and about which the guru claims to have an advantage of “*wisdom*” over them.

## 1.2 The Collaborative Discovery model

The *phenomenon* about which the guru and her followers interact is modeled by a *distribution*  $\mathcal{T}$  over  $\mathcal{X}^\Gamma$ , where  $\Gamma$  is the set of properties manifested by the phenomenon and  $\mathcal{X}$  is an arbitrary input space. The two input spaces mentioned in this paper are the binary space  $\mathcal{X} = \{0, 1\}$  and the finite categorical space  $\mathcal{X} = \{0, \dots, n\}$ . In the context of childhood development,  $\Gamma$  is the set of developmental milestones like “*first smile*”, and each follower (associated, for simplicity, with a parent of a single child) is represented by a sample  $u \in \mathcal{X}^\Gamma$  that describes the ages at which that child achieved each milestone. By time  $t$ , the follower discloses to the guru  $u|_{\Gamma_t}$ , the restriction of her sample  $u$  to a subset  $\Gamma_t \subseteq \Gamma$ . Additional attributes of  $u$  may be revealed later in time, others might be disclosed by the follower if prompted to do so, while certain attributes will remain forever latent.

The follower seeks the guru’s assistance in predicting “meaningful information” that is currently unknown to the follower. The guru and follower *interact* over a number of rounds but the follower will terminate the interaction if the guru is judged to be unhelpful (in a manner formalized below). During each round of interaction, the guru makes a prediction by announcing a distribution  $P_{\gamma_t}$  over  $\mathcal{X}$  that she claims is the true one for a latent attribute  $\gamma_t \notin \Gamma_t$ ; the follower has a distribution  $Q_{\gamma_t}$  that she believes corresponds to  $\gamma_t$ . (Modern gurus and followers are comfortable discussing probabilities rather than predicting a single event as is the case with pre-election polling results.) The way  $\gamma_t$  is selected from  $\Gamma \setminus \Gamma_t$  and its effect on the process is left to future work. The follower now queries  $\gamma_t$  and reports the true value, denoted  $u_{\gamma_t}$ , which is derived from Nature’s “true” distribution. After each round the follower updates the strength of her *retention* by the guru. We assume this strength is given by a *retention parameter*  $r_t$  that starts with a fixed value  $r_0$  and varies with time; once  $r_t$  turns negative the follower will be said to have lost all faith in the guru and therefore terminate the interaction. The main objective of the guru is to maintain  $r_t \geq 0$  for all  $t \geq 0$ ; jumping ahead, a distribution  $\mathcal{T}$  for which there exists a guru that, in expectation, manages to retain followers to eternity (or until  $t = |\Gamma|$  for finite  $\Gamma$ ) will be said to be  *$r_0$ -retainable* and the *retention complexity* of  $\mathcal{T}$  will be the minimal  $r_0$  such that  $\mathcal{T}$  is  $r_0$ -retainable (see Definitions 7, 8).

When the user updates her retention parameter at the end of round  $t$ , she uses a function  $\mathcal{S}(\cdot, \cdot, \cdot)$  that is real-valued and takes three inputs: (i) the guru’s predicted distribution  $P_{\gamma_t}$ ; (ii) the follower’s assessment of that distribution  $Q_{\gamma_t}$ ; and (iii) the value  $u_{\gamma_t}$  that materialized, sampled by Nature. The retention parameter at time  $t$  is given by

$$r_t = r_{t-1} + \mathcal{S}(P_{\gamma_t}, Q_{\gamma_t}, u_{\gamma_t}) \quad (1)$$

► **Remark (Simplifying assumptions).** The formula (1) makes the following assumptions on the follower’s update rule: that it is Markovian, uses  $r_{t-1}$  additively and does not depend on the follower’s identity nor on the identity of the attribute  $\gamma_t$  being predicted. Such assumptions are common when modeling human behavior and we leave the study of more general update functions to future work.

### 1.3 Retentive Scoring Rules

The definition of the function  $\mathcal{S}$  above, and the surprising corollaries of this definition, are what dominates the first part of our study. We assume  $\mathcal{S}$  belongs to a class of functions that *elicit* the true beliefs of both guru and follower regarding the distribution for the attribute  $\gamma_t$ . Truth eliciting rules are ones that incentivize (rational) players to supply the rule with what they believe to be the truth. A famous early example of a truth eliciting rule is that of a one-party *proper scoring rule*, which will be tightly related to our two-party retentive scoring rule  $\mathcal{S}$ , so we start with the simpler, one-party, case.

#### Proper (one-party) Scoring Rules

One-party *proper scoring rules* are used to compensate a single forecaster of Nature in a truth-eliciting manner; these rules are studied extensively in the Decision Theory literature [17, 23, 11] and have interesting connections to the fields of estimation, information theory, and machine learning; see [8] for a recent survey. A scoring rule receives a single forecast, which is a distribution  $P$  over  $\mathcal{X}$  as an input (say, this could be the temperature at noon tomorrow at a fixed location), and scores the forecaster based on the outcome selected by Nature (the actual temperature). A scoring rule is called *proper* if it is maximized by forecasting the true distribution. We recite the definition as it appears in [23, 11]:

► **Definition 1** (Proper Scoring Rule). Let  $\mathcal{P}$  be a convex set of distributions over an arbitrary input space  $\mathcal{X}$ . A (one-party) *scoring rule* is a function  $s : \mathcal{P} \times \mathcal{X} \rightarrow \mathbb{R}$ . The scoring rule  $s$  is *proper* with respect to  $\mathcal{P}$  if, for all  $R \in \mathcal{P}$  (viewed as Nature's true distribution), the expected score  $\mathbb{E}_{x \sim R} [s(P, x)]$  is maximized over  $P \in \mathcal{P}$  at  $P = R$ :

$$\forall P \in \mathcal{P} \quad \mathbb{E}_{x \sim R} [s(P, x)] \leq \mathbb{E}_{x \sim R} [s(R, x)] \quad (2)$$

Intuitively, when the agent forecasts a distribution  $P \in \mathcal{P}$  and event  $x \in \mathcal{X}$  materializes, the reward for the expert is  $s(P, x)$ . To increase clarity when one-party and two-party (retentive) scoring rules are involved, we will use a lowercase  $s$  to denote a proper (one-party) scoring rule, and a calligraphic  $\mathcal{S}$  to denote a retentive (two-party) one.

Many proper scoring rules can be constructed using elementary functions, for example the *logarithmic scoring rule*:

$$s(P, i) = \log p_i \quad (3)$$

and *Brier's scoring rule* [6]:

$$s(P, i) = 2p_i - \sum_j p_j^2 = 2p_i - \|P\|_2^2 \quad (4)$$

#### Retentive (Two-Party) Scoring Rules

In the spirit of proper scoring rules, we define a *retentive* scoring rule which involves two parties: guru and follower. Using an axiomatic approach, we start by defining the desired properties of such a rule:

► **Definition 2** (Retentive Scoring Rule). Let  $\mathcal{P}$  be a convex set of distributions over an arbitrary input space  $\mathcal{X}$ . A function  $\mathcal{S} : \mathcal{P} \times \mathcal{P} \times \mathcal{X} \rightarrow \mathbb{R}$  is called a *retentive scoring rule* if it satisfies the following conditions:

1. *Cost of ignorance*: For all distributions  $P \in \mathcal{P}$  and outcomes  $x \in \mathcal{X}$ ,

$$\mathcal{S}(P, P, x) = -1 \quad (5)$$

2. *Proper scorings*: for any distribution  $R \in \mathcal{P}$  dictated by Nature:

- a. *Guru-side*: For any fixed follower belief  $Q \in \mathcal{P}$ , the best guru prediction  $P \in \mathcal{P}$  is Nature's:

$$\mathbb{E}_{x \sim R} [\mathcal{S}(P, Q, x)] \leq \mathbb{E}_{x \sim R} [\mathcal{S}(R, Q, x)] \quad (6)$$

- b. *Follower-side*: For any fixed guru prediction  $P \in \mathcal{P}$ , the best follower belief  $Q \in \mathcal{P}$  is Nature's:

$$\mathbb{E}_{x \sim R} [\mathcal{S}(P, Q, x)] \geq \mathbb{E}_{x \sim R} [\mathcal{S}(P, R, x)] \quad (7)$$

Intuitively, the *cost of ignorance* condition models the “attention economy” cost of interaction, and captures the intuition that the follower will penalize gurus that are no “smarter” than he is. For example, no guru/meteorologist will get followers by predicting “100% chance of sun in the Sahara desert”. The predictions must be surprising to the followers. In the formal definition, the penalty constant is normalized to  $-1$  to simplify analysis.

The output of  $\mathcal{S}$  is a quantity that the guru wishes to maximize, because doing so will mean the follower is retained longer, as seen by Equation (1). Therefore, the *guru-side properness* requirement (Equation (6)) implies that a rational guru will strive to report the correct distribution used by Nature ( $R$ ), if the guru knows that distribution. In other words, we require the scoring rule to elicit *truthful* guru-side inputs.

Similarly, since the follower has a limited attention span, she is incentivized to judge the guru's quality “honestly”, and this is modeled by the *follower-side properness* condition (Equation (7)); it means the follower too will supply the rule  $\mathcal{S}$  with Nature's distribution, if known to her. Notice that the combination of the cost-of-ignorance and two properness results mean that a rational guru will not offer “obvious advice” about which both guru and follower “know the (same) truth”.

### Retentive Rule Construction

One-party scoring rules give rise to two-party retentive scoring rules in a straightforward way: Score the guru and follower *separately* based on Nature's outcome using, perhaps, two different functions, and define the retentive score as the difference between the one-party scores minus a fixed constant (to account for the cost-of-ignorance (5)). A retentive scoring rule of this form is said to be *separable*, and a special case is that of a *symmetric* rule, in which both guru and follower are scored using the same (one-party) scoring rule, formally:

► **Definition 3** (Symmetric Retentive Scoring Rule). A retentive scoring rule  $\mathcal{S} : \mathcal{P} \times \mathcal{P} \times \mathcal{X} \rightarrow \mathbb{R}$  is called *symmetric* if there exists a proper one-party scoring rule  $s : \mathcal{P} \times \mathcal{X} \rightarrow \mathbb{R}$  such that:

$$\mathcal{S}(P, Q, i) = s(P, i) - s(Q, i) - 1 \quad (8)$$

### Characterization

Restricting the discussion to *categorical distributions*, i.e., to cases where  $\mathcal{X}$  is finite, and assuming the retentive scoring rules are *analytic*, meaning that a uniformly convergent power series expansion exists about any  $P \in \mathcal{P}$ , our first main result is the following statement:

► **Theorem 4** (Retentive Scoring Rules are Symmetric). *The function  $\mathcal{S} : \mathcal{P} \times \mathcal{P} \times \mathcal{X} \rightarrow \mathbb{R}$  is an analytic retentive scoring rule for categorical distributions if and only if there exists a proper and analytic scoring rule  $s : \mathcal{P} \times \mathcal{X} \rightarrow \mathbb{R}$  such that:*

$$\mathcal{S}(P, Q, x) = s(P, x) - s(Q, x) - 1 \quad (9)$$

We find the statement somewhat surprising because it is not intuitively clear that a two-party retentive rule must be separable; symmetry follows rather directly from separability and the cost-of-ignorance assumption. For the proof – given in Section 2 – we use a known result which relates proper scoring rules to convex functions over the probability simplex. We show that each retentive scoring rule corresponds to a solution of a system of partial differential equations (PDEs). Solving the system and characterizing the family of solutions yields the result (see Subsection 2.2).

## 1.4 Memory Span

To model the different prediction capacities of gurus and followers, the forecasting abilities of both types of agents in the Collaborative Discovery model are characterized by a parameter called *memory span*, defined below.

A variety of psychological studies could be summarized by saying that the human short-term memory has a capacity of about “seven, plus-or-minus two” *chunks*, where each chunk can be roughly defined as a collection of elementary information relating to a single concept [18, 24]. What counts as a chunk depends on the knowledge of the person being tested. For instance, a word is a single chunk for a speaker of the language but is many chunks for someone who is totally unfamiliar with the language and sees the word as a collection of phonetic segments.

In the world of child development, young parents (who usually don’t have significant experience or formal child-development education) are likely to predict that their child will start walking around the average time for the entire population. Child development experts, on the other hand, usually have better ability to pick up subtle developmental signals from observed child behavior, and provide a better prediction based on them.

In this spirit, we proceed with the formal definition. In what follows, let  $\Delta(\mathcal{X}^\Gamma)$  denote the simplex of probability distributions over  $\mathcal{X}^\Gamma$  and  $\Delta(\mathcal{X})$  is the simplex of distributions over  $\mathcal{X}$ .

► **Definition 5** (Memory Span). Let  $\mathcal{T} \in \Delta(\mathcal{X}^\Gamma)$  be a distribution. An agent is said to have *memory span*  $m \geq 0$  when its prediction  $P_\gamma \in \Delta(\mathcal{X})$  for coordinate  $\gamma_t \in \Gamma$  of an instance  $u \in \mathcal{X}^\Gamma$  with disclosed parameters  $\Gamma_t \subseteq \Gamma$  (i.e. for which  $u_{|\Gamma_t}$  is known) is based on  $m$  disclosed coordinates or less:

$$\forall \gamma \in \Gamma, \exists I_t \subseteq \Gamma_t : |I_t| \leq m, P_\gamma = (\mathcal{T}_\gamma \mid u_{|I_t}) \quad (10)$$

where  $(\mathcal{T}_\gamma \mid u_{|I_t})$  is the marginal distribution of  $\mathcal{T}$  on coordinate  $\gamma$ , conditioned on the event that the coordinates  $I_t$  are set to  $u_{|I_t}$ .

Intuitively, this means that every prediction of an agent is based on its entire knowledge of at most  $m$  coordinates. When  $m = 0$ , a prediction is only based on the marginal distribution of the corresponding parameter in the entire population.

### 1.4.1 Monotonicity

Our second result, stated next, says that if guru  $G$  is “smarter” than guru  $G'$ , meaning her memory span ( $m_g$ ) is greater than his ( $m'_g$ ), the smarter guru  $G$  will also have higher success in retaining followers, in expectation. (Whether this optimistic result holds in the real world is highly debatable.) This result is not implied directly by the definition of the Collaborative Discovery model, and shows that our model exhibits intuitive and desirable properties that substantiate its theoretical appeal:

► **Theorem 6** (Knowledgeable Gurus Retain Better). *Let  $\mathcal{S} : \mathcal{P} \times \mathcal{P} \times \mathcal{X} \rightarrow \mathbb{R}$  be an analytic retentive scoring rule, let  $G_1, G_2$  be two gurus with memory spans  $m_g^{(1)} \geq m_g^{(2)}$ . Then for any distribution  $\mathcal{T}$ , any coordinate  $x$ , and follower with memory span  $m_f \leq m_g^{(2)}$ :*

$$\mathbb{E}_{\mathcal{T}}[\mathcal{S}(P_1, Q, x)] \geq \mathbb{E}_{\mathcal{T}}[\mathcal{S}(P_2, Q, x)] \quad (11)$$

where  $P_1, P_2 \in \Delta(\mathcal{X})$  are the distributional forecasts of gurus  $G_1$  and  $G_2$  respectively, and  $Q \in \Delta(\mathcal{X})$  is the belief of the follower.

A technical discussion of the theorem and its proof are provided in Section 3.

### 1.4.2 Retainability as a Function of Memory Span Discrepancy

From here on we assume that the guru has memory span  $m_g$ , and her follower has memory span  $m_f$  and moreover, both parties provide to the retentive scoring rule a distribution that is the correct marginal  $\mathcal{T}_{\gamma_t} \mid u_{\downarrow J_t}$ , conditioned on some subset of  $J_t \subset \Gamma_t$  of size  $m_g$  for the guru and  $m_f$  for the follower, respectively. Under this assumption, notice that if  $m_g = m_f$  then both parties supply the same distribution, so the cost-of-ignorance assumption of Definition 2 means the follower will terminate the interaction within  $r_0$  steps; in other words, ignorant gurus will not prevail. Henceforth assume  $m_g > m_f$ . Combining the concepts of limited user attention, retentive scoring, and limited memory span, we can now ask: Is it possible for the guru to retain her follower throughout the process? This leads to the concept of *retainability*:

► **Definition 7** (Retentively Learnable Distribution). Let  $\mathcal{T} \in \Delta(\mathcal{X}^\Gamma)$ , and assume  $|\Gamma| = n$ . Given a retentive scoring rule  $\mathcal{S} : \mathcal{P} \times \mathcal{P} \times \mathcal{X} \rightarrow \mathbb{R}$ , guru memory span  $m_g \geq 0$ , follower memory span  $m_f \geq 0$ , and an initial retention parameter  $r_0 > 0$ , we say that  $\mathcal{T}$  is *retentively learnable* with respect to  $(\mathcal{S}, m_g, m_f, r_0)$  if there exists an ordering  $\gamma_1, \dots, \gamma_n$  of  $\Gamma$ , and a sequence of sets  $I_1, \dots, I_n$  such for all  $t \in [n]$ :

1.  $I_t \subseteq \{\gamma_1, \dots, \gamma_{t-1}\}$
2.  $|I_t| \leq m_g$
3. For every sequence of sets  $J_1, \dots, J_n$  such that  $J_t \subseteq \{\gamma_1, \dots, \gamma_{t-1}\}$ ,  $|J_t| \leq m_f$ :

$$r_t = r_0 + \sum_{t'=1}^t \mathcal{S}((\mathcal{T}_{\gamma_{t'}} \mid u_{\downarrow I_{t'}}), (\mathcal{T}_{\gamma_{t'}} \mid u_{\downarrow J_{t'}}), \mathcal{T}) \geq 0 \quad (12)$$

Intuitively, a probability distribution is retentively learnable when it is possible to maintain a positive retention parameter throughout the process. From (12) we can see that increasing  $r_0$  does not hurt retainability. In other words, for  $r'_0 > r_0$ , if a distribution is retentively learnable with respect to  $(\mathcal{S}, m_g, m_f, r_0)$ , then it is also retentively learnable for  $(\mathcal{S}, m_g, m_f, r'_0)$ . We know that attention is a very limited resource, so we cannot expect it to be arbitrarily large. This leads to the following question: How large should the “initial retention” be in order for the guru to sustain her follower throughout the collaborative discovery process?



► **Definition 8** (Retention Complexity). The *retention complexity* of a distribution  $\mathcal{T} \in \Delta(\mathcal{X}^\Gamma)$  with respect to  $(\mathcal{S}, m_g, m_f)$  is the minimal value of  $r_0$  such that  $\mathcal{T}$  is retainable:

$$r_{\mathcal{S}, m_g, m_f}(\mathcal{T}) = \min \{r_0 \mid \mathcal{T} \text{ is retainable with respect to } (\mathcal{S}, m_g, m_f, r_0)\} \quad (13)$$

The discussion above leads to the following notion of retentively learnable families of distributions and properties. Recall that a property  $P$  over alphabet  $\mathcal{X}$  is a set of finite strings over  $\mathcal{X}$ . Let  $P_n = P \cap \mathcal{X}^n$  denote the set of strings in  $P$  of length  $n$ . Let  $\mathbb{N}_P = \{n \in \mathbb{N} \mid P_n \neq \emptyset\}$  be the set of lengths of strings that belong to  $P$ . The *family of distributions induced by  $P$*  is the family of uniform distributions over  $P_n$ , defined for  $n \in \mathbb{N}_P$ .

► **Definition 9** (Retentively Learnable Distributions and Properties). Let  $\mathbb{D} = \{\mathcal{T}_i\}_{i \in I}$  be a family of categorical distributions, where each  $\mathcal{T}_i$  is supported on strings of length  $n_i$  over alphabet  $\mathcal{X}$  (the set  $I$  may be infinite). Let  $\mathcal{S} : \mathcal{P} \times \mathcal{P} \times \mathcal{X} \rightarrow \mathbb{R}$  be a scoring rule as in Definition 3. We say  $\mathbb{D}$  is *retentively learnable* with respect to  $\mathcal{S}$  if there exist constants  $r_0$  and  $m_g > m_f$  such that each  $\mathcal{T}_i \in \mathbb{D}$  is retainable with respect to  $(\mathcal{S}, m_g, m_f, r_0)$ .

Similarly, a property  $P \subset \mathcal{X}^*$  is said to be *retentively learnable* with respect to  $\mathcal{S}$  if the family of distributions induced by  $P$  is retentively learnable.

Our main complexity result is the following. Recall that the property of linear functions over  $\mathbb{F}_2$  is the set of functions  $f : \mathbb{F}_2^k \rightarrow \mathbb{F}_2$  that satisfy  $f(x+y) = f(x) + f(y)$  for all  $x, y \in \mathbb{F}_2^k$ . For this property we have  $\mathbb{N}_P = \{n = 2^k \mid k \in \mathbb{N}\}$ .

► **Theorem 10** (Linear functions are retentively learnable). *The property of linear functions over the two-element field  $\mathbb{F}_2$  is retentively learnable.*

We elaborate on this result, and related ones, next.

## 1.5 The Retention Complexity of Linear Spaces

To initiate the study of the retentive learning within the context of computational complexity, a natural starting point is that of uniform distributions over linear spaces; linearity is the first object of study in other notable fields of complexity, like property testing [5, 13] and PAC learning [25].

Consider a realization of the model in which each attribute ranges over a binary space, i.e.,  $\mathcal{X}^\Gamma = \{0, 1\}^n$ . The Binary Attributes model describes a universe where each attribute is either present or not for a given user.

We start by redefining the problem using finite-field linear algebra, and then study the retention complexity of several natural families of linear codes, including the Walsh-Hadamard codes and the family of random Low Density Parity Check (LDPC) codes.

In particular, identify  $\{0, 1\}$  with the two-element finite field  $\mathbb{F}_2$  and consider a uniform distribution  $\mathcal{U}$  over a linear space  $U \subseteq \mathbb{F}_2^n$ . Let  $U^\perp$  denote the space dual to  $U$ . Let  $d(U)$  denote the Hamming distance of  $U$  (and  $d(U^\perp)$  is its dual distance), recalling that distance is equal to the minimum Hamming weight of a non-zero word in  $U$  (or  $U^\perp$ , respectively). We assume the guru has infinite memory span and the follower has memory span 0. (The study of the general case of  $0 < m_f < m_g < \infty$  is left for future work.) This means the follower's distribution for each  $i \in [n]$  is the uniform distribution on  $\mathbb{F}_2$  (this assumes  $U$  is not constant on any  $i \in [n]$ ).

We shall use a retentive scoring rule denoted  $\mathcal{S}_{\text{bin}}$ , that has expected value 1 when the guru can predict the next coordinate exactly, i.e., when the value of that coordinate depends linearly on the values of coordinates exposed thus far, and has expected value  $-1$  otherwise, when the distribution on that coordinate is linearly independent of all previously revealed bits.



The following result sets the bounds for our study of retention complexity in this setting, establishing a connection between the retention complexity of  $\mathcal{U}$  and its algebraic properties:

► **Lemma 11** (Retention Complexity Bounds for Linear Spaces). *For a uniform distribution  $\mathcal{U}$  over a linear space  $U \subseteq \mathbb{F}_n^2$  with unbounded guru memory span and zero follower memory span, the retention complexity satisfies:*

$$d(U^\perp) - 1 \leq r_{(\mathcal{S}_{\text{bin}}, \infty, 0)}(\mathcal{U}) \leq \dim(U) \quad (14)$$

Next, we show that the both bounds are tight. We begin by showing that a uniform distribution over codewords of the *Walsh-Hadamard* (WH) code achieves the lower retention complexity bound:

► **Lemma 12** (Walsh-Hadamard Retention Complexity). *For all  $k \in \mathbb{N}$ , a  $k$ -dimensional Walsh-Hadamard code satisfies:*

$$r_{(\mathcal{S}_{\text{bin}}, 2, 0)}(\text{WH}) = 2 \quad (15)$$

As the  $n$ -dimensional Walsh-Hadamard code represents the set of linear functions over  $\mathbb{F}_n^2$ , proving this lemma will also imply Theorem 10.

Finally, we show that a random LDPC code achieves the upper bound (up to multiplicative constants) with high probability, implying that collaborative discovery can be very hard on arbitrary linear spaces:

► **Theorem 13** (LDPC Retention Complexity). *For a proper choice of constants  $c, d > 0$  and sufficiently large  $n$ , the retention complexity of a random  $(c, d)$ -regular LDPC code over  $\mathbb{F}_2^n$  is linear with high probability:*

$$r_{(\mathcal{S}_{\text{bin}}, \infty, 0)}(\text{LDPC}) \underset{\text{w.h.p.}}{=} \Omega(\dim(\text{LDPC})) \quad (16)$$

The proofs of these results are provided in Subsection 4.2. The most technically challenging one is the third one and relies on the lower bounds for the testability of random LDPC codes of [4].

## 1.6 Related work

The study of reputation systems is interested in ranking gurus in “meaningful” ways, and is highly investigated empirically and theoretically; cf. [20, 21] and references therein. Closest in rigour to our approach are the papers by (i) Ban and Linial [2] which uses the theory of random processes to identify situations where gurus (called “experts” there) can be robustly ranked, assuming user participation continues indefinitely, and (ii) Chan et al. [7] that classifies interactive crowd-computation games using a small list of modeling parameters.

In the context of machine learning, the task of detecting users who are likely to stop participating in a voluntary system is known as *churn prediction*. For this task, machine learning algorithms are trained to recognize typical usage patterns and predict the likelihood of a termination [26, 9]. Even though general machine learning models provide good “black-box” churn predictors when trained correctly, gaining deep understanding of the underlying phenomena might be challenging.

Comparing our model to prior work, there are two main differences. First, our aim is to model the dynamics of *long-term* interaction between a follower and her guru about a *single complex phenomenon* of interest, asking when do followers abandon their gurus. Second, we

are interested in the *mathematical properties of phenomena* that are prone to collaborative discovery, meaning that for these phenomena a “good” guru will successfully instruct her followers from start to end without losing their attention and faith. This motivation is somewhat similar to that taken in the field of Property Testing [13] which attempts to understand which properties are amenable to “testing”.

## 1.7 Discussion of main contributions and future directions

The properties of retentive scoring rules, the effect of memory span discrepancy on the retention of followers, and the study of retention complexity of specific distributions are the main topics of this work.

We point out a few questions that emerge from the paper:

1. The gurus and followers studied here are assumed to have optimal knowledge of the distribution, up to their memory span limit. In particular, a guru with infinite memory span does not need to *learn* the distribution at all. However, in most realistic settings the distribution is unknown, leading to the question of learning distributions in a way that also maintains good retention properties. For instance, suppose the distribution is an *unknown* linear space  $U$  with retention complexity  $r$ . What is the minimal number of followers with initial retention parameter  $r_0 > r$  (say,  $r_0 = 2 \cdot r$ ) that will be “spent” or “lost” by the guru before she learns enough about  $U$  to fully retain new followers? This particular question is highly relevant to automated gurus that seek to attract users while maintaining high reputation (e.g., high app-store ratings).
2. The gurus and followers used here are computationally unbounded (or, more precisely, bounded only by attention span). Realistically, the computational complexity of computing marginals and evaluating which new attribute  $\gamma_t$  to interact about will be highly non-trivial.
3. Walsh-Hadamard codes are locally testable, correctable and decodable, while random LDPC codes have none of these properties; moreover, the retention complexity for both families of codes is approximately equal to their query complexity (for testability and correctability). This leads to our first question: *Are there tighter connections between retention complexity and query complexity of locally testable/correctable codes?* Do all  $q$ -query locally testable (or correctable) codes have retention complexity  $f(q)$  for some function that depends only on  $q$  and is independent of  $n$  (input size)? Likewise, it seems interesting to ask whether retention complexity is related to basic machine learning measures like VC dimension.

## 2 Retentive Scoring

In this section we study retentive scoring rules, and prove Theorem 4.

### 2.1 Preliminaries and Notations

#### Categorical Probability Distributions

Recall that a *categorical distribution* is a discrete probability distribution that describes the possible results of a random event that can take one of  $K$  possible outcomes. In this section, we assume  $\mathcal{P}$  is a convex set of categorical distributions with  $K = (n + 1)$  possible outcomes, i.e.  $\mathcal{X} = \{0, \dots, n\}$ . We define the number of possible outcomes as  $n + 1$  instead of  $n$  to simplify later calculations.

In addition, recall that the space of categorical distributions with  $(n + 1)$  possible outcomes is equivalent to the  $n$ -dimensional simplex:

$$\mathcal{P} \subseteq \Delta^n = \left\{ (p_0, \dots, p_n) \in \mathbb{R}^{n+1} \mid \sum_i p_i = 1; \forall i : p_i \geq 0 \right\} \quad (17)$$

where  $p_i$  is the probability of categorical event  $i$ .

### Expected Score Notation

Recall Definition 2. Following the conventions of the proper scoring literature, and given probability distributions  $P, Q, R \in \mathcal{P}$ , we denote the *expected retentive score* as:

$$\mathcal{S}(P, Q, R) \equiv \mathbb{E}_{i \sim R} [\mathcal{S}(P, Q, i)] \quad (18)$$

To avoid difficulties in (18), we will assume  $\mathcal{S}(P, Q, R)$  exists and is finite. Similarly, for one-party scoring rules, the common notation of *expected score* is:

$$s(P, R) \equiv \mathbb{E}_{i \sim R} [s(P, i)] \quad (19)$$

The analysis below will use both the single event notation  $\mathcal{S}(P, Q, i)$  and the expected score notation  $\mathcal{S}(P, Q, R)$  (and similarly for one-party scoring rules). To avoid confusion, we will always use upper-case letters to denote random variables and lower-case letters to denote events.

► **Remark (Scoring Rules on Infinite Sample Spaces).** Similar to proper (one-party) scoring rules, it is possible to define retentive scoring rules on infinite sample spaces using measure-theoretic tools. Computers are finite, and therefore many applications can be modeled as finite-dimensional categorical distributions. In this work we consider the finite sample space for concreteness and simplicity, and leave the rigorous measure-theoretic analysis to future work.

### Characterization of Proper Scoring Rules

One of the fundamental results in the research of proper scoring rules is the characterization theorem, first stated by [17], which defines a correspondence between proper scoring rules and convex functions over the probability simplex. We start with some preliminary definitions, and proceed with the characterization theorem itself:

► **Definition 14 (Subgradient).** A function  $\nabla^* G : \mathcal{P} \rightarrow \mathbb{R}^{n+1}$  is a subgradient of  $G$  at the point  $P$  if the following inequality holds for all  $Q \in \mathcal{P}$ :

$$G(Q) \geq G(P) + \langle \nabla^* G(P), (Q - P) \rangle \quad (20)$$

where  $\langle \cdot, \cdot \rangle$  denotes the euclidean inner product over  $\mathbb{R}^{n+1}$ :  $\langle X, Y \rangle = \sum_{i=0}^n x_i y_i$ .

► **Remark (Subgradients of Differentiable Functions).** If  $G$  is differentiable at  $P \in \mathcal{P}$  then  $G$  has a *unique* subgradient at  $P$  and it equals the gradient  $\nabla G = \left( \frac{\partial G}{\partial p_0}, \dots, \frac{\partial G}{\partial p_n} \right)$  at  $P$ .

Recall that a real-valued function  $G : \mathcal{P} \rightarrow \mathbb{R}$  is convex if:  $G((1 - \lambda)P + \lambda Q) \leq (1 - \lambda)G(P) + \lambda G(Q)$  for all  $P, Q \in \mathcal{P}$  and  $\lambda \in [0, 1]$ .

► **Lemma 15** ([15], Theorem 2.1).  $G : \mathcal{P} \rightarrow \mathbb{R}$  is convex if and only if it has a subgradient  $\nabla^* G$  at each point  $P \in \mathcal{P}$ .

► **Theorem 16** (McCarthy's Theorem, [11]). *A scoring rule  $s : \mathcal{P} \times \Omega \rightarrow \mathbb{R}$  is proper relative to  $\mathcal{P}$  if and only if there exists a convex, real-valued function  $G$  on  $\mathcal{P}$  such that:*

$$s(P, i) = G(P) - \langle \nabla^* G(P), P \rangle - (\nabla^* G)_i \quad (21)$$

where  $(\nabla^* G)_i$  is the  $i$ th component of  $(\nabla^* G)$ .

We also define the *Generalized Entropy* as the convex function which is induced by the proper scoring rule:

► **Definition 17** (Generalized Entropy). The convex function  $G(P) = s(P, P)$  induced by a proper scoring rule  $s$  is called the *generalized entropy function* of  $s$ .

Note that a convex general entropy function exists for every proper scoring rule by Theorem 16. For the logarithmic scoring rule defined in (3), the associated general entropy function is the additive inverse of the Shannon entropy:  $G(P) = \sum_{i=0}^n p_i \log p_i$ . Additional information-theoretic quantities can be generalized using proper scoring rules. See [8] for a recent review.

## 2.2 Separability of Retentive Scoring Rules

In this section, we prove that every proper retentive scoring rule can be written as the difference between two proper scoring rules. Recall Theorem 4:

► **Theorem 4** (Retentive Scoring Rules are Symmetric). *The function  $\mathcal{S} : \mathcal{P} \times \mathcal{P} \times \mathcal{X} \rightarrow \mathbb{R}$  is an analytic retentive scoring rule for categorical distributions if and only if there exists a proper and analytic scoring rule  $s : \mathcal{P} \times \mathcal{X} \rightarrow \mathbb{R}$  such that:*

$$\mathcal{S}(P, Q, x) = s(P, x) - s(Q, x) - 1 \quad (9)$$

The proof has several steps:

1. We verify that symmetric retentive scoring rules are indeed proper (Lemma 18).
2. Conversely, we first define the notion of a *separable scoring rule*, which is a rule which can be written as the difference between two one-party scoring rules. Given a retentive scoring rule, we use the proper scoring characterization theorem (Theorem 16) to construct a system of partial differential equations which describes the constraints that must be satisfied by such a rule (Lemma 20). We then solve the characterizing system of partial differential equations (Lemma 21), and show that every possible solution corresponds to a separable retentive scoring rule (Lemma 22).
3. Finally, we show that every separable retentive scoring rule with constant cost of ignorance is also symmetric, proving the theorem.

We proceed by stating and proving the lemmas, and conclude the section by proving the theorem itself.

### Preliminaries

The proofs of Lemma 18 and Lemma 20 rely on the formalism of proper scoring rules and retentive scoring rules. The proof of Claim 21 relies on basic results from the theory of quasi-linear partial differential equations (refer to [19] for a thorough introduction). For  $D \subseteq \mathbb{R}^n$ , we will refer to a function  $f : D \rightarrow \mathbb{R}$  as *analytic* if its Taylor expansion about  $\mathbf{x} \in D$  converges to  $f(\mathbf{x})$  for all  $\mathbf{x} \in D$ . We use  $e_i \in \mathbb{R}^n$  to denote the  $i$ th vector of the standard basis. The gradient of a differentiable function  $g(\mathbf{x}, \mathbf{y}) : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$  with respect to  $\mathbf{x} \in \mathbb{R}^n$  is denoted by  $\frac{\partial g}{\partial \mathbf{x}} \equiv \left( \frac{\partial g}{\partial x_1}, \dots, \frac{\partial g}{\partial x_n} \right)^T$ .

### 2.2.1 Symmetric Rules are Retentive

► **Lemma 18.** *Let  $\mathcal{S} : \mathcal{P} \times \mathcal{P} \times \mathcal{X} \rightarrow \mathbb{R}$  be a scoring rule. If there is a proper scoring rule  $s : \mathcal{P} \times \mathcal{X} \rightarrow \mathbb{R}$  such that:  $\mathcal{S}(P, Q, i) = s(P, i) - s(Q, i) - 1$ , then  $\mathcal{S}(P, Q, i)$  is retentive.*

**Proof.** Let  $P, Q, R \in \mathcal{P}$ . Using (8), the expected score  $\mathcal{S}(P, Q, R)$  is:

$$\mathcal{S}(P, Q, R) = s(P, R) - s(Q, R) - 1 \quad (22)$$

$s$  is proper, and therefore  $s(P, R) \leq s(R, R)$ . Plugging into (22) we obtain:

$$\mathcal{S}(P, Q, R) \leq s_1(R, R) - s_2(Q, R) = \mathcal{S}(R, Q, R) \quad (23)$$

satisfying (6). Similarly,  $s_2$  is also proper, and therefore:

$$\mathcal{S}(P, Q, R) \geq s_1(P, R) - s_2(R, R) = \mathcal{S}(P, R, R) \quad (24)$$

satisfying (6). For  $Q = P$  we get  $\mathcal{S}(P, P, i) = -1$  for all  $i \in \mathcal{X}$ , and therefore  $\mathcal{S}$  is retentive according to Definition 2. ◀

### 2.2.2 Retentive Rules are Separable

We start by formally defining the notion of a separable scoring rule:

► **Definition 19** (Separable Retentive Scoring Rule). A proper retentive scoring rule  $\mathcal{S} : \mathcal{P} \times \mathcal{P} \times \Omega \rightarrow \mathbb{R}$  is called *separable* if there exists two proper scoring rules  $s_1, s_2 : \mathcal{P} \times \Omega \rightarrow \mathbb{R}$  such that:

$$\mathcal{S}(P, Q, i) = s_1(P, i) - s_2(Q, i) \quad (25)$$

We also say that a two-party scoring rule is *proper* if it satisfies (6), (7). In the following lemma, we say that a bi-variate function  $G : \mathcal{P} \times \mathcal{P} \rightarrow \mathbb{R}$  is convex with respect to its first argument if  $G(P, Q)$  is a convex function of  $P$  for any constant  $Q \in \mathcal{P}$ ; convexity with respect to the second argument is similarly defined by switching the roles of  $P$  and  $Q$ .

► **Lemma 20** (Characterization by subgradients). *A two-party scoring rule  $\mathcal{S}$  is proper with respect to class  $\mathcal{P}$  if and only if there exist two functions  $G, H : \mathcal{P} \times \mathcal{P} \rightarrow \mathbb{R}$  such that:*

1.  $G(P, Q)$  is convex with respect to  $P$ .
2.  $H(P, Q)$  is convex with respect to  $Q$ .
3. For all  $P, Q, R \in \mathcal{P}$ :

$$G + \langle \nabla_P^* G, (R - P) \rangle = -(H + \langle \nabla_Q^* H, (R - Q) \rangle) \quad (26)$$

where  $\nabla_P^* G$  is a subgradient of  $G(P, Q)$  with respect to its first argument, and  $\nabla_Q^* H$  is a subgradient of  $H(P, Q)$  with respect to its second argument.

**Proof.** For the first direction, let  $\mathcal{S}(P, Q, i)$  be a proper retentive scoring rule, and define  $s_Q(P, i) \equiv \mathcal{S}(P, Q, i)$ . Using (6) we obtain that  $s_Q(P, R) \leq s_Q(R, R)$ . Therefore  $s_Q$  is proper, and according to Theorem 16 there exists a convex function  $G_Q : \mathcal{P} \rightarrow \mathbb{R}$  that depends on  $Q$ , such that:

$$s_Q(P, i) = G_Q(P) - \langle \nabla^* G_Q(P), P \rangle + (\nabla^* G_Q(P))_i \quad (27)$$

## 12:14 The Complexity of User Retention

where  $(\nabla^* G_Q(P))_i$  is  $i$ th entry of  $\nabla^* G_Q$  at point  $P$ . Similarly, define  $s_P(Q, i) \equiv \mathcal{S}(P, Q, i)$ . By the same reasoning and using (7) we obtain that  $-s_P$  is proper, and therefore there exists a convex function  $H_P : \mathcal{P} \rightarrow \mathbb{R}$  such that:

$$-s_P(Q, i) = H_P(Q) - \langle \nabla^* H_P(Q), Q \rangle + (\nabla^* H_P(Q))_i \quad (28)$$

Define  $G(P, Q) \equiv G_Q(P)$  and  $H(P, Q) \equiv H_P(Q)$ . Note that  $G$  is convex with respect to  $P$  and  $H$  is convex with respect to  $Q$ , satisfying conditions 1, 2. Let  $R \in \mathcal{P}$ . Using the fact that  $s_P(P, R) = s_Q(Q, R)$ , we can combine (27), (28) to obtain:

$$G + \langle \nabla_P^* G, (R - P) \rangle = -(H + \langle \nabla_Q^* H, (R - Q) \rangle) \quad (29)$$

satisfying condition 3.

Conversely, let  $G, H$  be the functions which satisfy the three conditions above. Define:

$$s_Q(P, i) \equiv G - \langle \nabla_P^* G, P \rangle + (\nabla_P^* G)_i \quad (30)$$

$$s_P(Q, i) \equiv -\left(H - \langle \nabla_Q^* H, H \rangle + (\nabla_Q^* H)_i\right) \quad (31)$$

Note that  $s_Q = -s_P$  by equation (26), and that  $s_P$  and  $-s_Q$  are proper by Theorem 16.

Define  $\mathcal{S}(P, Q, i) = s_Q(P, i) = -s_P(Q, i)$ .  $s_Q$  is proper, and therefore  $\mathcal{S}(P, Q, R) \leq \mathcal{S}(R, Q, R)$ , satisfying the properness condition in (6). Similarly, the properness of  $-s_P$  implies  $\mathcal{S}(P, R, R) \leq \mathcal{S}(P, Q, R)$ , satisfying (7), and therefore  $\mathcal{S}$  is proper.  $\blacktriangleleft$

The following lemma contains a solution of a partial differential equation that will assist us in solving the characterizing equations of proper retentive scoring rules. We obtain the solution using basic tools from the theory of partial differential equations, and the proof is given in Appendix A for completeness:

► **Claim 21.** *Let  $D \subseteq \mathbb{R}^n$  such that  $\mathbf{x}, \mathbf{y} \in D$ . For every analytic function  $u : D \times D \rightarrow \mathbb{R}$  satisfying the equation*

$$u(\mathbf{x}, \mathbf{y}) - \sum_{i=1}^n (y_i - x_i) \frac{\partial u(\mathbf{x}, \mathbf{y})}{\partial x_i} = 0 \quad (32)$$

*there exist functions  $\alpha_1, \dots, \alpha_n : D \rightarrow \mathbb{R}$  such that:*

$$u(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n \alpha_i(\mathbf{y})(y_i - x_i) \quad (33)$$

The following lemma is the heart of this part of the proof of Theorem 4 .

► **Lemma 22** (Proper Retentive Rules are Separable). *Let  $\mathcal{S} : \mathcal{P} \times \mathcal{P} \times \mathcal{X} \rightarrow \mathbb{R}$  be a retentive scoring rule. If  $\mathcal{S}$  is a proper retentive scoring rule with analytic generalized entropy functions, then there exist two functions  $s_1, s_2 : \mathcal{P} \times \Omega \rightarrow \mathbb{R}$  such that  $\mathcal{S}(P, Q, i) = s_1(P, i) - s_2(Q, i)$ .*

Proof outline:

1. Given a proper retentive scoring rule, Lemma 20 implies the existence of two generalized entropy functions  $G, H : \mathcal{P} \times \mathcal{P} \rightarrow \mathbb{R}$  related by equation (26).
2. We choose a parametrization for points on the simplex  $\Delta^n$ , and use it to define (26) in the convex domain  $D = \{(x_1, \dots, x_n) \in \mathbb{R}_+^n \mid \sum_i x_i \leq 1\}$ .
3. We simplify the resulting equation, and solve it using Claim 21.

4. Applying the correspondence established in Theorem 16 between convex functions on the simplex and proper scoring rules, we show that the generalized entropy functions  $G, H$  induce a separable scoring rule.

Following the conventions of multivariate calculus, in the proof we will use the  $\cdot$  symbol to denote the euclidean inner product over  $\mathbb{R}^n$ :  $\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^n x_i y_i$ . In addition, the proof employs terms from the theory of multivariate convex analysis: Given a non-empty convex subset  $S \subseteq \mathbb{R}^n$ , its *affine hull*  $\text{Aff}(S)$  is the smallest affine set containing  $S$ . A *relative interior point* is a member of the set  $\{x \in S : \exists \epsilon > 0, N_\epsilon(x) \cap \text{Aff}(S) \subseteq S\}$ , where  $N_\epsilon(x)$  is the  $\epsilon$ -ball around point  $x$ . Refer to [22] for an introduction to convex analysis. In the proof, we also use the *gradient theorem* for line integrals, which is a common generalization of the fundamental theorem of calculus. We recall it here without proof. Refer to Wikipedia entry [14] for discussion and proof:

► **Claim 23** (Gradient Theorem). *Let  $\varphi : U \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$  and  $\gamma$  is any curve from  $\mathbf{p}$  to  $\mathbf{q}$ . Then:*

$$\varphi(\mathbf{q}) - \varphi(\mathbf{p}) = \int_{\gamma[\mathbf{p}, \mathbf{q}]} \nabla \varphi(\mathbf{r}) \cdot d\mathbf{r} \quad (34)$$

**Proof of Lemma 22.** Let  $\mathcal{S}$  be a proper retentive scoring rule. By Lemma 20, there exist two functions  $G, H : \mathcal{P} \times \mathcal{P} \rightarrow \mathbb{R}$  such that  $G(P, Q)$  is convex with respect to its first argument,  $H(P, Q)$  is convex with respect to its second argument, and equation (26) is satisfied.

When  $P, Q$  and  $R$  are categorical random variables with  $n+1$  possible outcomes, equation (26) is defined over the  $n$ -dimensional simplex  $\Delta^n$ . Let  $D = \{(x_1, \dots, x_n) \in \mathbb{R}_+^n \mid \sum_i x_i \leq 1\}$ . Each point  $P$  on the simplex can be represented by a vector  $P = (p_0, \dots, p_n) \in \mathbb{R}_+^{n+1}$  such that  $\sum_{i=0}^n p_i = 1$ . To simplify the constraints, we define a bijection  $f : \Delta^n \rightarrow D$  as follows:

$$f(P) \equiv (p_1, \dots, p_n) \in \mathbb{R}^n \quad (35)$$

$$f^{-1}(\mathbf{x}) \equiv \left(1 - \sum_{i=1}^n x_i, x_1, \dots, x_n\right) \in \Delta^n \quad (36)$$

using this bijection, we represent each point on the simplex using a  $n$ -dimensional vector in the domain. Denote:  $P \equiv f^{-1}(\mathbf{x})$ ,  $Q \equiv f^{-1}(\mathbf{y})$ ,  $R \equiv f^{-1}(\mathbf{z})$ ,  $f(\mathcal{P}) \equiv \{f(P) \mid P \in \mathcal{P}\}$ .

Using this correspondence, we also define  $g(\mathbf{x}, \mathbf{y}) \equiv G(P, Q)$ ,  $h(\mathbf{x}, \mathbf{y}) \equiv H(P, Q)$ . The assumption that  $G, H$  are analytic implies that the gradients of each function coincide with their corresponding subgradients (See Remark 2.1).

We will now write (26) using the new parametrization. Let  $\mathbf{x}, \mathbf{y}, \mathbf{z} \in f(\mathcal{P})$ . For the left-hand side of (26) we obtain:

$$\frac{\partial g}{\partial \mathbf{x}} \cdot (\mathbf{z} - \mathbf{x}) = \sum_{i=1}^n \frac{\partial g}{\partial x_i} (z_i - x_i) \quad (37)$$

[Calculate the derivative of  $g$  using the chain rule]

$$= \sum_{i=1}^n \left( \frac{\partial G}{\partial p_i} - \frac{\partial G}{\partial p_0} \right) (z_i - x_i) \quad (38)$$

[Rearrange the summations]

$$= \frac{\partial G}{\partial p_0} \sum_{i=1}^n (-z_i + x_i) + \sum_{i=1}^n \frac{\partial G}{\partial p_i} \cdot (z_i - x_i) \quad (39)$$



## 12:16 The Complexity of User Retention

$$= \frac{\partial G}{\partial p_0} \left( \left( 1 - \sum_{i=1}^n z_i \right) - \left( 1 - \sum_{i=1}^n x_i \right) \right) + \sum_{i=1}^n \frac{\partial G}{\partial p_i} \cdot (z_i - x_i) \quad (40)$$

[Use the definition of  $\mathbf{x}, \mathbf{z}$ ]

$$= \sum_{i=0}^n \frac{\partial G}{\partial p_i} (r_i - p_i) \quad (41)$$

$$= \nabla G \cdot (R - P) \quad (42)$$

A similar argument on the right-hand side of (26) shows that  $\nabla H \cdot (R - Q) = h + \frac{\partial h}{\partial \mathbf{y}} \cdot (\mathbf{z} - \mathbf{y})$ , and therefore the system defined in (26) is equivalent to:

$$\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in f(\mathcal{P}) : g + \frac{\partial g}{\partial \mathbf{x}} \cdot (\mathbf{z} - \mathbf{x}) = - \left( h + \frac{\partial h}{\partial \mathbf{y}} \cdot (\mathbf{z} - \mathbf{y}) \right) \quad (43)$$

We will now simplify (43) using its linear properties. Denote the affine hull of  $f(\mathcal{P})$  by  $\text{Aff}(f(\mathcal{P})) \equiv \mathbf{v}_0 + V$ , and assume  $\mathbf{v}_0$  is a relative interior point. Taking  $\mathbf{z} = \mathbf{v}_0$  in equation (43) yields:

$$g + \frac{\partial g}{\partial \mathbf{x}} \cdot (\mathbf{v}_0 - \mathbf{x}) = - \left( h + \frac{\partial h}{\partial \mathbf{y}} \cdot (\mathbf{v}_0 - \mathbf{y}) \right) \quad (44)$$

Similarly, denote the  $i$ th basis vector of  $V$  by  $\bar{\mathbf{v}}_i$ . For any  $i \in [\dim V]$ , taking  $\mathbf{z} = \mathbf{v}_0 + \bar{\mathbf{v}}_i$  in equation (43), with appropriate scaling of  $\bar{\mathbf{v}}_i$  such that  $\mathbf{z} \in f(\mathcal{P})$ , yields:

$$\forall i \in [\dim V] : g + \frac{\partial g}{\partial \mathbf{x}} \cdot (\mathbf{v}_0 + \bar{\mathbf{v}}_i - \mathbf{x}) = - \left( h + \frac{\partial h}{\partial \mathbf{y}} \cdot (\mathbf{v}_0 + \bar{\mathbf{v}}_i - \mathbf{y}) \right) \quad (45)$$

Subtracting (44) from (45) we obtain:

$$\forall i \in [\dim V] : \frac{\partial g(\mathbf{x}, \mathbf{y})}{\partial \mathbf{x}} \cdot \bar{\mathbf{v}}_i = - \frac{\partial h(\mathbf{x}, \mathbf{y})}{\partial \mathbf{y}} \cdot \bar{\mathbf{v}}_i \quad (46)$$

thus  $\frac{\partial g(\mathbf{x}, \mathbf{y})}{\partial \mathbf{x}}$  and  $-\frac{\partial h(\mathbf{x}, \mathbf{y})}{\partial \mathbf{y}}$  are equal component-wise, and therefore  $\frac{\partial g(\mathbf{x}, \mathbf{y})}{\partial \mathbf{x}} + \frac{\partial h(\mathbf{x}, \mathbf{y})}{\partial \mathbf{y}}$  is orthogonal to the affine hull:

$$\forall \mathbf{v} \in V : \left( \frac{\partial g(\mathbf{x}, \mathbf{y})}{\partial \mathbf{x}} + \frac{\partial h(\mathbf{x}, \mathbf{y})}{\partial \mathbf{y}} \right) \cdot \mathbf{v} = 0 \quad (47)$$

Note that  $(\mathbf{z} - \mathbf{x}), (\mathbf{z} - \mathbf{y}), (\mathbf{y} - \mathbf{x}) \in V$ . Substitute (47) back into (43) to obtain:

$$g + \frac{\partial g}{\partial \mathbf{x}} \cdot (\mathbf{y} - \mathbf{x}) = -h \quad (48)$$

Apply  $\frac{\partial}{\partial \mathbf{y}}$  on both sides to get:

$$\frac{\partial g}{\partial \mathbf{y}} + \frac{\partial}{\partial \mathbf{y}} \sum_{i=1}^n \frac{\partial g}{\partial x_i} (y_i - x_i) = - \frac{\partial h}{\partial \mathbf{y}} \quad (49)$$

And using (47) again we obtain:

$$\frac{\partial}{\partial \mathbf{y}} \sum_{i=1}^n \frac{\partial g}{\partial x_i} (y_i - x_i) = 0 \quad (50)$$

Which is equivalent to:

$$\forall k \in [n] : \frac{\partial g}{\partial y_k} + \sum_{i=1}^n (y_i - x_i) \frac{\partial}{\partial x_i} \frac{\partial g}{\partial y_k} = 0 \quad (51)$$

This is a system of  $n$  independent first-order partial differential equations for each element in  $\frac{\partial g}{\partial \mathbf{y}}$ . Using Claim 21, we obtain the general solution for each  $k$ :

$$\forall k \in [n], \exists \alpha_{k,1}, \dots, \alpha_{k,n} : \frac{\partial g}{\partial y_k} = \sum_{i=1}^n \alpha_{k,i}(\mathbf{y})(y_i - x_i) \quad (52)$$

Packing back the equations to vector form, we define a matrix operator  $A : D \rightarrow \mathbb{R}^{n \times n}$  such that  $A_{i,j}[\mathbf{y}] = \alpha_{k,i}(\mathbf{y})$ . The system in (52) can now be compactly represented using matrix multiplication:

$$\frac{\partial g}{\partial \mathbf{y}} = A[\mathbf{y}] (\mathbf{y} - \mathbf{x}) \quad (53)$$

We now use the correspondence established in Theorem 16 to show that the generalized entropy functions  $G, H$  induce a separable scoring rule. Applying the gradient theorem (34) along the curve  $\gamma(t) = \mathbf{0} + t\mathbf{y}$  for  $t \in [0, 1]$  yields:

$$g(\mathbf{x}, \mathbf{y}) - g(\mathbf{x}, \mathbf{0}) = \int_0^1 \left( \mathbf{y}^T \left( \frac{\partial g}{\partial \mathbf{y}} \Big|_{\mathbf{x}, t\mathbf{y}} \right) \right) dt \quad (54)$$

$$= \int_0^1 (\mathbf{y}^T A[t\mathbf{y}] (t\mathbf{y} - \mathbf{x})) dt \quad (55)$$

Denote  $\psi(\mathbf{x}) \equiv g(\mathbf{x}, \mathbf{0})$  and  $\varphi(\mathbf{x}, \mathbf{y}) \equiv \int_0^1 (\mathbf{y}^T A[t\mathbf{y}] (t\mathbf{y} - \mathbf{x})) dt$ . Note that  $\varphi(\mathbf{x}, \mathbf{y})$  is a linear function of  $\mathbf{x}$ . The scoring rule which corresponds to  $g$  is given by Theorem 16:

$$S(\mathbf{x}, \mathbf{y}, \mathbf{z}) = g(\mathbf{x}, \mathbf{y}) + \frac{\partial g(\mathbf{x}, \mathbf{y})}{\partial \mathbf{x}} \cdot (\mathbf{z} - \mathbf{x}) \quad (56)$$

$$= \underbrace{\psi(\mathbf{x}) + \frac{\partial \psi(\mathbf{x})}{\partial \mathbf{x}} \cdot (\mathbf{z} - \mathbf{x})}_{\equiv s_1} + \underbrace{\varphi(\mathbf{x}, \mathbf{y}) + \frac{\partial \varphi(\mathbf{x}, \mathbf{y})}{\partial \mathbf{x}} \cdot (\mathbf{z} - \mathbf{x})}_{\equiv s_2} \quad (57)$$

The terms denoted by  $s_1$  only depend on  $\mathbf{x}$  and  $\mathbf{z}$ , and therefore  $s_1 = s_1(\mathbf{x}, \mathbf{z})$ . In addition,  $\varphi(\mathbf{x}, \mathbf{y})$  is a linear function of  $\mathbf{x}$  and therefore both  $\frac{\partial \varphi(\mathbf{x}, \mathbf{y})}{\partial \mathbf{x}}$  and  $\left( \varphi(\mathbf{x}, \mathbf{y}) - \frac{\partial \varphi(\mathbf{x}, \mathbf{y})}{\partial \mathbf{x}} \cdot \mathbf{x} \right)$  do not depend on  $\mathbf{x}$ , thus  $s_2 = s_2(\mathbf{y}, \mathbf{z})$ . The scoring rule  $S(\mathbf{x}, \mathbf{y}, \mathbf{z})$  can therefore be written in the following form:

$$S(\mathbf{x}, \mathbf{y}, \mathbf{z}) = s_1(\mathbf{x}, \mathbf{z}) - s_2(\mathbf{y}, \mathbf{z}) \quad (58)$$

and applying the reverse transformation from  $\mathbf{x}, \mathbf{y}, \mathbf{z} \in D$  to  $P, Q, R \in \mathcal{P}$  implies the separability of the original scoring rule  $\mathcal{S}$ .  $\blacktriangleleft$

### 2.2.3 Concluding the Proof

We can now conclude the section by proving the characterization theorem for retentive scoring rules. For the final proof, recall Definition 3 of symmetric retentive rules.

**Proof of Theorem 4.** Given a proper scoring rule  $s : \mathcal{P} \times \mathcal{X} \rightarrow \mathbb{R}$  such that  $\mathcal{S}(P, Q, i) = s(P, i) - s(Q, i) - 1$ , we can apply Lemma 18 to show that  $\mathcal{S}(P, Q, i)$  is retentive. Conversely, given an analytic retentive scoring rule, we can apply Lemma 22 and obtain  $s_1, s_2$  such that  $\mathcal{S}(P, Q, i) = s_1(P, i) - s_2(Q, i)$ . The rule  $\mathcal{S}$  is retentive, and therefore satisfies (5). for all  $P \in \mathcal{P}$  and  $Q = P$  we obtain:

$$\mathcal{S}(P, P, i) = s_1(P, i) - s_2(P, i) = -1 \quad (59)$$

and therefore  $s_1(P, i) = s_2(P, i) - 1$  for all  $P$ , proving that  $\mathcal{S}$  is symmetric.  $\blacktriangleleft$

### 3 Monotonicity

In this section we show that expected retention score in each round grows with the size of memory span, proving Theorem 6:

► **Theorem 6 (Knowledgeable Gurus Retain Better).** *Let  $\mathcal{S} : \mathcal{P} \times \mathcal{P} \times \mathcal{X} \rightarrow \mathbb{R}$  be an analytic retentive scoring rule, let  $G_1, G_2$  be two gurus with memory spans  $m_g^{(1)} \geq m_g^{(2)}$ . Then for any distribution  $\mathcal{T}$ , any coordinate  $x$ , and follower with memory span  $m_f \leq m_g^{(2)}$ :*

$$\mathbb{E}_{\mathcal{T}} [\mathcal{S}(P_1, Q, x)] \geq \mathbb{E}_{\mathcal{T}} [\mathcal{S}(P_2, Q, x)] \quad (11)$$

where  $P_1, P_2 \in \Delta(\mathcal{X})$  are the distributional forecasts of gurus  $G_1$  and  $G_2$  respectively, and  $Q \in \Delta(\mathcal{X})$  is the belief of the follower.

The proof will require a definition and a lemma: We first define the notion of *Localized Expected Gain* (Definition 24), which is a set function that quantifies the expected score for different choices of prior data. We then show that this function is monotone by proving Lemma 25, and use the result to prove the theorem itself.

#### Preliminaries

We denote the jointly distributed vector by  $(X_1, \dots, X_n) \sim \mathcal{T}$ . The marginal distribution of coordinate  $i$  is denoted by  $X_i$ . For  $t \in [n]$  and  $I \subseteq [n]$  such as  $t \notin I$ , the marginal value of coordinate  $t$  conditioned on the event  $X_{\setminus I} = x_{\setminus I}$  is denoted by  $(X_t \mid x_I)$ . When probability calculations are involved, we will omit the harpoon notation for brevity, and  $x_I$  and  $x_{\setminus I}$  will be used interchangeably.

► **Definition 24 (Localized Expected Gain).** Let  $(X_1, \dots, X_t) \sim D \in \Delta(\mathcal{X}^t)$  be a set of  $t$  jointly-distributed random variables, let  $I \subseteq [t-1]$ , and let  $s : \Delta(\mathcal{X}) \times \mathcal{X} \rightarrow \mathbb{R}$  be a proper (one-party) scoring rule. The *localized expected gain* is a set function  $f : 2^{[t-1]} \rightarrow \mathbb{R}$  defined as follows:

$$\forall I \subseteq [t-1] : f(I) \equiv \mathbb{E}_{(x_1, \dots, x_t) \sim D} [s((X_t \mid x_I), x_t)] \quad (60)$$

Intuitively, the localized expected gain function  $f(I)$  describes the expected score when  $X_{\setminus I}$  is being used as a prior. For example, for the log scoring rule  $s(P, i) = \log p_i$  defined in (3), the associated expected localized gain function is:

$$f_{\log}(I) = \sum_{x_I} \Pr(x_I) \sum_{x_t} \Pr(x_t \mid x_I) \log \Pr(x_t \mid x_I) = -H(X_t \mid X_I) \quad (61)$$

which is the additive inverse of the conditional entropy of  $X_t$  given  $X_I$ .

We now show that this function is also monotone for general proper scoring rules, which means that expected scores don't decrease when adding prior information, or "more knowledge doesn't hurt" regardless of the proper scoring rule being used:

► **Lemma 25.**  *$f$  is a monotone set function:*

$$\forall I \subseteq J \subseteq [t-1] : f(I) \leq f(J) \quad (62)$$

**Proof.** We start with the definition of  $f(I)$ :

$$f(I) = \mathbb{E}_{(x_1, \dots, x_t) \sim D} [s((X_t | x_I), x_t)] \quad (63)$$

$$= \sum_{x_{[t-1]}, x_t} \Pr(x_{[t-1]}, x_t) s((X_t | x_I), x_t) \quad (64)$$

[Decompose  $\Pr(x_{[t-1]}, x_t)$  using the law of total probability]

$$= \sum_{x_{[t-1]}, x_t} \Pr(x_J) \Pr(x_t | x_J) \Pr(x_{[t-1] \setminus J} | x_t, x_J) s((X_t | x_I), x_t) \quad (65)$$

[ $s$  does not depend on  $y_{[t] \setminus J}$ . Rearrange the summation]

$$= \sum_{x_J} \Pr(x_J) \sum_{x_t} \Pr(x_t | x_J) s((X_t | x_I), x_t) \sum_{x_{[t-1] \setminus J}} \Pr(x_{[t-1] \setminus J} | x_t, x_J) \quad (66)$$

[The rightmost factor is equal to 1]

$$= \sum_{x_J} \Pr(x_J) \sum_{x_t} \Pr(x_t | x_J) s((X_t | x_I), x_t) \quad (67)$$

Using the definition of expected one-party score defined in (19), we obtain that the rightmost factor in (67) is the expected score of  $P = (X_t | x_I)$  when the reference distribution is  $R = (X_t | x_J)$ :

$$f(I) = \sum_{x_J} \Pr(x_J) s((X_t | x_I), (X_t | x_J)) \quad (68)$$

We can now use the properness of  $s$  (see Definition 1) to obtain:

$$f(I) \leq \sum_{x_J} \Pr(x_J) s((X_t | x_J), (X_t | x_J)) \quad (69)$$

and apply steps (63),..., (67) in reverse order to obtain

$$\sum_{x_J} \Pr(x_J) s((X_t | x_J), (X_t | x_J)) = f(J) \quad (70)$$

proving that  $f(I) \leq f(J)$ . ◀

Using Lemma 25 we can generalize the result to retentive scoring rules, and prove the monotonicity theorem for retentive scoring rules:

**Proof of Theorem 6.** Guru 1 has memory span  $m_g^1$ , and therefore  $P_1 = (\mathcal{T} | u_{\downarrow I_1})$  such that  $|I_1| = m_g^1$ . Similarly, for guru 2 we obtain  $P_2 = (\mathcal{T} | u_{\downarrow I_2})$  such that  $|I_2| = m_g^2$  and for the follower  $Q = (\mathcal{T} | u_{\downarrow J})$  such that  $|J| = m_f$ .

$\mathcal{S}$  is analytic, and therefore symmetric according to Theorem 4. Denote  $\mathcal{S}(P, Q, i) = s(P, i) - s(Q, i) - 1$ . Taking the expectation over  $\mathcal{T}$  we obtain:

$$\mathbb{E}_{\mathcal{T}} [\mathcal{S}(P, Q, i)] = \mathbb{E}_{\mathcal{T}} [s(P, i)] - \mathbb{E}_{\mathcal{T}} [s(Q, i)] - 1 \quad (71)$$

Using Definition 24 we obtain:

$$\mathbb{E}_{\mathcal{T}} [\mathcal{S}(P, Q, i)] = f(I) - f(J) - 1 \quad (72)$$

## 12:20 The Complexity of User Retention

When  $m_g^1 \geq m_g^2$  and under the optimal choice of  $I_1$ , there exists  $I'_1$  such that  $I_2 \subseteq I'_1$  and  $f(I'_1) \leq f(I_1)$ . Applying Lemma 25 we obtain:

$$\mathbb{E}_{\mathcal{T}}[\mathcal{S}(P_1, Q, i)] = f(I_1) - f(J) - 1 \quad (73)$$

$$\geq f(I'_1) - f(J) - 1 \quad (74)$$

$$\geq f(I_2) - f(J) - 1 \quad (75)$$

$$= \mathbb{E}_{\mathcal{T}}[\mathcal{S}(P_2, Q, i)] \quad (76)$$

and therefore  $\mathbb{E}_{\mathcal{T}}[\mathcal{S}(P_1, Q, i)] \geq \mathbb{E}_{\mathcal{T}}[\mathcal{S}(P_2, Q, i)]$ . ◀

### 4 The Binary Attributes Model

Under the Binary Attributes model, the universe of users is modeled using a  $k$ -dimensional linear subspace of  $\mathbb{F}_2^n$ .

$$U = \text{span}\{\bar{u}_1, \dots, \bar{u}_k\} \quad (77)$$

where  $\bar{u}_1, \dots, \bar{u}_k \in \mathbb{F}_2^n$  are a choice of basis vectors for the subspace. Under this realization of the Collaborative Discovery model, each user is represented using an  $n$ -dimensional binary vector, formally  $\mathcal{X}^n = \mathbb{F}_2^n$ .

#### Preliminaries

This section will assume familiarity with basic linear algebra over finite fields. A *view*  $I \subseteq [n]$  of a vector  $u \in \mathbb{F}_2^n$ , denoted by  $u|_I$ , is a linear projection of  $u$  to the subspace  $V_I = \text{span}\{e_i \mid i \in I\}$ . Similar to the previous section, we omit the harpoon notation when complex conditional probability expressions are involved. Given a vector space  $U$ , its *dual space* is defined as the set of linear constraints:  $U^\perp \equiv \{v \in \mathbb{F}_2^n \mid \forall u \in U : \langle u, v \rangle = 0\}$ . The *support* of a vector  $u \in U$  is the of coordinates that contain non-zero elements:  $\text{support}(u) = \{i \mid u_i \neq 0\}$ . We denote the hamming distance of a vector  $u \in U$  by  $d(u) = |\text{support}(u)|$ . The hamming distance of the space  $U$  is defined as  $d(U) = \min_{u \in U \setminus \{0\}} d(u)$ .

#### 4.1 User Types as a Linear Subspace

We follow with a rigorous definition of the process under the Binary Attributes realization:

##### Initialization

At the start of the Collaborative Discovery process, the type of user  $u$  is picked uniformly from  $U$ , all the coordinates are undisclosed, and the initial retention parameter is  $r_0$ . We will denote the uniform random variable over the linear space by  $\mathcal{U} \sim \text{Uniform}(U)$ .

##### Prediction Rounds

During each round, the expert picks a coordinate  $i$  and provides a prediction distribution  $P \in \Delta(\{0, 1\})$  for its value. The retentive scoring function for this realization of the model is:

$$\mathcal{S}_{\text{bin}}(P, Q, x) = 2 \log_2 p_x - 2 \log_2 q_x - 1 \quad (78)$$

where  $x \in \{0, 1\}$ .  $\mathcal{S}_{\text{bin}}$  can be represented as  $\mathcal{S}_{\text{bin}}(P, Q, x) = s(Q, x) - s(P, x) - 1$ , where  $s(P, x) = 2 \log_2 p_x$  is the logarithmic scoring rule defined in (3), and therefore  $\mathcal{S}_{\text{bin}}$  is symmetric according to Definition 3.

We'll proceed to show that  $\mathcal{S}_{\text{bin}}$  has very intuitive properties. To do that, we start with a few basic claims about the structure of this probability space. The claims can be proved using basic linear algebra and probability. Proofs are included in Appendix B:

► **Claim 26.** *Let  $I \subseteq [n]$ . For every vector  $u_I \in U_I$ :*

$$\Pr(\mathcal{U}_I = u_I) = 2^{-\dim(U_I)} \quad (79)$$

For the following claim, recall that a *singleton distribution* is a probability distribution in which a single outcome has probability 1.

► **Claim 27.** *Let  $I \subseteq [n]$  and  $m \in [n] \setminus I$ , and assume a vector  $u \in \mathbb{F}_2^n$  has been picked uniformly at random from a vector space  $U$ .  $\Pr(u_m \mid u_I)$  is a singleton distribution if and only if  $e_m \in U_{[n] \setminus I}^\perp$ .*

► **Claim 28.** *Let  $U$  be a linear space over  $\mathbb{F}_2^n$ , and let  $I \subseteq [n], m \in [n] \setminus I$ .  $e_m \in U_{[n] \setminus I}^\perp$  if and only if  $\dim(U_{[I]}) = \dim(U_{[I] \cup \{m\}})$ .*

Using this framework, we now have enough tools to characterize the dynamics of scoring rule we defined:

► **Lemma 29** (Binary Attributes Scoring Rule Dynamics). *For a uniform distribution  $\mathcal{U}$  over a linear space  $U$  without constant bits, the retention score for a collaborative discovery process with infinite expert locality and zero layperson locality is given by:*

$$\mathcal{S}_{\text{bin}}((X_m \mid x_I), X_m, \mathcal{U}) = \begin{cases} 1 & e_m \in U_{[n] \setminus I}^\perp \\ -1 & \text{otherwise} \end{cases} \quad (80)$$

$$= \begin{cases} 1 & \dim(U_{[I] \cup \{m\}}) = \dim(U_{[I]}) \\ -1 & \dim(U_{[I] \cup \{m\}}) = \dim(U_{[I]}) + 1 \end{cases} \quad (81)$$

**Proof.** When  $e_m \notin U_{[n] \setminus I}^\perp$ , we get that  $\dim(U_{[I] \cup \{m\}}) = 1$ , allowing us to apply Claim 26 and obtain  $\Pr(u_m = 0 \mid u_I) = \frac{1}{2}$ .

When  $e_m \in U_{[n] \setminus I}^\perp$  there exists  $v \in U^\perp, I' \subseteq I$  such that  $\text{support}(v) = I' \cup \{m\}$ . Claim 27 implies that  $u_m$  is determined given  $u_{I'}$ .

Combining the results, we obtain for all  $I \subseteq [n], m \notin I$ :

$$\Pr(u_m = 0 \mid u_I) \in \begin{cases} \{0, 1\} & e_m \in U_{[n] \setminus I}^\perp \\ \{\frac{1}{2}\} & \text{otherwise} \end{cases} \quad (82)$$

There are no constant bits in  $U$ , and therefore  $\dim U_{[I] \cup \{m\}} = 1$  for all  $m \in [n]$ . By Claim 26 we obtain that the marginal distribution for each coordinate is uniform, and therefore a layperson with zero locality will always predict a uniform distribution.

Plugging (82) into the definition of  $\mathcal{S}_{\text{bin}}$  in equation (78), the score for the first case is  $\log_2 \frac{1}{2 \cdot 0.5^2} = 1$ , and the score for the second case is  $\log_2 \frac{0.5^2}{2 \cdot 0.5^2} = -1$ , leading to equation (80). The transition from (80) to (81) is given by Claim 28. ◀

## 4.2 Retention Complexity of Linear Codes

We will now apply the notion of retention complexity introduced in Definition 8 to the Binary Attributes model. We will first show that there exist non-trivial upper and lower bounds for retention complexity in this realization of the Collaborative Discovery model, and then show that the bounds are tight. Recall Lemma 11:

► **Lemma 30** (Retention Complexity Bounds for Linear Spaces). *For a uniform distribution  $\mathcal{U}$  over a linear space  $U \subseteq \mathbb{F}_n^2$  with unbounded guru memory span and zero follower memory span, the retention complexity satisfies:*

$$d(U^\perp) - 1 \leq r_{(\mathcal{S}_{\text{bin}}, \infty, 0)}(\mathcal{U}) \leq \dim(U) \quad (14)$$

**Proof of Lemma 11.** The retention parameter at the end of each round  $t$  is defined according to equation (1):

$$r_t = r_0 + \sum_{i=1}^t \mathcal{S}_{\text{bin}}((X_{\sigma_i} \mid x_{I_i}), X_{\sigma_i}, \mathcal{U}) \quad (83)$$

For the lower bound, observe that  $U^\perp|_{[n] \setminus I_t}$  does not contain any singleton element when  $|I_t| \leq d(U^\perp) - 2$ . Since  $|I_t| \leq t - 1$  by definition, we can combine the inequalities and obtain that no punctured-dual-space singleton exists when  $t \leq d(U^\perp) - 1$ . We can now apply Lemma 29 and obtain that  $\mathcal{S}_{\text{bin}}((X_{\sigma_i} \mid x_{I_i}), X_{\sigma_i}, \mathcal{U}) = -1$  for all  $i \in \{1, \dots, t\}$ . Plugging into the retention parameter at time  $t = d(U^\perp) - 1$ :

$$r_t = r_0 + \sum_{i=1}^{d(U^\perp)-1} (-1) = r_0 - (d(U^\perp) - 1) \quad (84)$$

And the positivity constraint on  $r_t$  implies that  $r_0 \geq (d(U^\perp) - 1)$ .

For the upper bound, assume without loss of generality that the first  $k = \dim(U)$  coordinates of  $U$  are linearly independent, and set  $\sigma_i = i, I_i = \{1, \dots, (i-1)\}$  for all  $i \in \{1, \dots, k\}$ . Observe that:

$$\dim(U|_{I_i}) = \begin{cases} i-1 & 1 \leq i \leq k \\ k & k < i \end{cases} \quad (85)$$

Applying Lemma 29 we get:

$$\mathcal{S}_{\text{bin}}((X_{\sigma_i} \mid x_{I_i}), X_{\sigma_i}, \mathcal{U}) = \begin{cases} -1 & 1 \leq i \leq k \\ 1 & k < i \end{cases} \quad (86)$$

Hence for  $r_0 = k$  we get  $r_t \geq 0$  for all  $t \in \{1, \dots, n\}$ . ◀

In the asymptotic setting it is common to consider  $n, k \rightarrow \infty$ . In this case,  $d(U^\perp)$  can stay constant, forming a large gap between the bounds. We will proceed to show that the upper and lower bounds are indeed tight in the asymptotic setting.

### 4.2.1 Linear Functions are Retentively Learnable

Let  $n = 2^k - 1$ . Given a binary message  $x \in \{0, 1\}^k$ , the *Walsh-Hadamard* code (WH) encodes the message into a codeword  $\text{WH}(x)$  using an encoding function  $\text{WH} : \{0, 1\}^k \rightarrow \{0, 1\}^n$ , such that for every  $y \in (\{0, 1\}^k \setminus \{0^k\})$ , the  $y$ th coordinate of  $\text{WH}(x)$  is equal to  $(x \cdot y)$ .



Walsh-Hadamard is the space of linear functions over  $\mathbb{F}_2^n$ . Note that we slightly deviate from the common definition by omitting the 0th coordinate which is always equal to zero. It is a  $[2^k - 1, k, 2^{k-1}]_2$  locally-correctable code with  $q = 2$  queries. See [1] for a thorough discussion of Walsh-Hadamard codes and its applications in theoretical computer science.

We will show that a uniform distribution over the  $k$ -dimensional WH code achieves the retention complexity lower bound for all  $k \in \mathbb{N}$ . Recall Lemma 12:

► **Lemma 31** (Walsh-Hadamard Retention Complexity). *For all  $k \in \mathbb{N}$ , a  $k$ -dimensional Walsh-Hadamard code satisfies:*

$$r_{(\mathcal{S}_{\text{bin}}, 2, 0)}(\text{WH}) = 2 \quad (15)$$

In order to prove the lemma, we first characterize the constraints of the WH code (Claim 32, Claim 33), and then use the results to construct an explicit formula for the retention score when the  $\mathcal{S}_{\text{bin}}$  retentive score rule is being used (Lemma 34), giving an upper bound for  $r_{(\mathcal{S}_{\text{bin}}, 3, 0)}(\text{WH})$  which is equal to the lower bound we established in Lemma 11. Proofs for the claims can be found in Appendix B.

► **Claim 32.** *Let  $y^{(1)}, \dots, y^{(m)} \in \left(\{0, 1\}^k \setminus \{0^k\}\right)$ .*

$$\left(\sum_{i=1}^m e_{y^{(i)}}\right) \in \text{WH}^\perp \iff \sum_{i=1}^m y^{(i)} = 0 \quad (87)$$

► **Claim 33.**

$$d(\text{WH}^\perp) = 3 \quad (88)$$

► **Lemma 34.** *Let  $k > 0$ , and let  $u$  be a uniformly-sampled vector from the  $k$ -dimensional WH code. Set a natural ordering over the coordinates ( $\gamma_t = t$  for  $t \in \{1, \dots, 2^k - 1\}$ ), and set a sequence of subsets:*

$$I_t = \begin{cases} \{2^{\lfloor \log_2 t \rfloor}, t - 2^{\lfloor \log_2 t \rfloor}\} & t > 2, \text{ and } t \text{ is not a power of } 2 \\ \emptyset & \text{otherwise} \end{cases}$$

*For collaborative discovery with respect to  $\mathcal{S} = \mathcal{S}_{\text{bin}}$ ,  $m_g = 2$ ,  $m_f = 0$  and  $r_0 = 2$ , the ordering  $\gamma_t$  and sequence of sets  $I_t$  satisfies:*

$$r_t = t - 2^{\lfloor \log_2 t \rfloor} \quad (89)$$

*for all  $1 \leq t < 2^k$ .*

**Proof.** By induction:

For the base case  $t \in \{1, 2\}$ . According to the definition,  $\gamma_1 = 1, \gamma_2 = 2$ , and  $I_1 = I_2 = \emptyset$ . We use Claim 33 and an argument similar to the one in Lemma 11 to show that there's no singleton in the punctured dual-space in the first two rounds. The guess in the first two rounds will therefore be a uniform one, and  $r_1 = 1, r_2 = 0$ . Indeed we can substitute 0, 1 into (89) see that  $1 - 2^{\lfloor \log_2 1 \rfloor} = 1$  and  $1 - 2^{\lfloor \log_2 2 \rfloor} = 1$ .

For  $t > 2$ , assume the retention parameter formula holds for  $t - 1$ , and consider the two following cases:

- When  $t$  is not a power of two, it can be represented as the XOR between two preceding coordinates, for example  $t' = 2^{\lfloor \log_2 t \rfloor}$  and  $t'' = t - 2^{\lfloor \log_2 t \rfloor}$ . Note that  $I_t = \{t', t''\}$ , and that  $|I_t| = 2$ , satisfying the  $m_g = 2$  memory span constraint. Using Claim 32 we obtain that  $e_t$  is a singleton in the punctured dual-space, and therefore  $r_t = r_{t-1} + 1$  by Lemma 29. Using the induction hypothesis and the fact that  $\lfloor \log_2 t \rfloor = \lfloor \log_2 (t-1) \rfloor$  when  $t$  is not a power of two, we obtain:

$$\begin{aligned} r_t &= r_{t-1} + 1 \\ &= (t-1) - 2\lfloor \log_2 (t-1) \rfloor + 1 \\ &= t - 2\lfloor \log_2 t \rfloor \end{aligned}$$

- When  $t$  is a power of two, it cannot be represented as the XOR between preceding coordinates, as for all of them the index of the most significant bit is strictly less than  $\log_2 t$ . By Lemma 29 we obtain that  $r_t = r_{t-1} - 1$ , and using the fact that  $\lfloor \log_2 t \rfloor = \lfloor \log_2 (t-1) \rfloor + 1$  when  $t$  is a power of two we indeed get:

$$\begin{aligned} r_t &= r_{t-1} - 1 \\ &= (t-1) - 2\lfloor \log_2 (t-1) \rfloor - 1 \\ &= t - 2\lfloor \log_2 t \rfloor \end{aligned} \quad \blacktriangleleft$$

► **Remark (Non-punctured Walsh-Hadamard).** In the non-punctured Walsh-Hadamard code, the 0th coordinate is not omitted, and always equal to zero. Offsetting the sequences in Lemma 34 can show that the same upper bound also holds for the non-punctured version of the Walsh-Hadamard code.

Combining the results proves Lemma 12:

**Proof of Lemma 12.** Lemma 34 shows an upper bound of 2 for the retention complexity of WH. Lemma 11, together Claim 33, tells us that this is also the lower bound for the retention complexity in this case, and therefore  $r_{(\mathcal{S}_{\text{bin}}, 2, 0)}(\text{WH}) = 2$ .  $\blacktriangleleft$

We can now conclude and prove Theorem 10. Recall the theorem statement:

► **Theorem 10** (Linear functions are retentively learnable). *The property of linear functions over the two-element field  $\mathbb{F}_2$  is retentively learnable.*

**Proof of Theorem 10.** For linear functions over  $\mathbb{F}_2^n$ , the corresponding family of categorical distributions is  $\mathbb{D} = \{\mathcal{U}_n\}_{n=2^k}$ , where  $\mathcal{U}_i$  is the family of uniform distributions over the non-punctured Walsh-Hadamard code. Lemma 34 shows that each  $\mathcal{U}_i$  is retainable with respect to  $(\mathcal{S}_{\text{bin}}, 2, 0, 2)$ .  $\blacktriangleleft$

## 4.2.2 Random LDPC Codes are Asymptotically Hard to Retain

Let  $G = (L, R, E)$  be a bipartite multigraph with  $|L| = n$ ,  $|R| = m$ . Associate a distinct Boolean variable  $x_i$  with any  $i \in L$ . For each  $j \in R$ , let  $N(j) \subseteq L$  be the set of neighbors of  $j$ . The  $j$ th constraint is  $A_j(x_1, \dots, x_n) = \sum_{i \in N(j)} x_i \pmod{2}$ . The code defined by  $G$  is:

$$\mathcal{C}(G) = \{x \in \{0, 1\}^n \mid \forall j \in [m] : A_j(x) = 0\}$$

A random  $(c, d)$ -regular LDPC code of length  $n$  is obtained by taking  $\mathcal{C}(G)$  for a random  $(c, d)$ -regular  $G$  with  $n$  left vertices. Random LDPC codes were first described and analyzed by [10]. We will show that a randomly chosen LDPC code asymptotically achieves the upper bound for retention complexity with high probability. Recall Theorem 13:

► **Theorem 13** (LDPC Retention Complexity). *For a proper choice of constants  $c, d > 0$  and sufficiently large  $n$ , the retention complexity of a random  $(c, d)$ -regular LDPC code over  $\mathbb{F}_2^n$  is linear with high probability:*

$$r_{(\mathcal{S}_{\text{bin}}, \infty, 0)}(\text{LDPC}) \stackrel{\text{w.h.p.}}{=} \Omega(\dim(\text{LDPC})) \quad (16)$$

► **Definition 35** ( $(q, \mu)$  code locality, [4]). A linear space  $V$  is  $(q, \mu)$ -local if every  $v \in V$  that is a sum of at least  $\mu m$  basis vectors has  $d(v) \geq q$ .

The following lemma shows that a random LDPC code has  $(q, \mu)$ -locality with high probability for a proper choice of parameters:

► **Lemma 36** ([4], Lemma 3.6). *Fix odd integer  $c \geq 7$  and constants  $\mu, \delta, d > 0$  satisfying:*

$$\mu \leq \frac{c^{-2}}{100}; \quad \delta < \mu^c; \quad d > \frac{2\mu c^2}{(\mu^c - \delta)^2} \quad (90)$$

*Then, for all sufficiently large  $n$ , with high probability for a random  $(c, d)$ -regular graph  $G$  with  $n$  left vertices and  $m = \frac{c}{d}n$  right vertices, the corresponding LDPC code  $\mathcal{C}(G)$  is linearly-independent, and  $(\delta n, \mu)$ -local.*

► **Remark 37** (A Proper Choice of Parameters). For our proof of Theorem 13, the constants in (90) need be chosen such that  $\delta - \frac{2\mu c}{d} \geq 0$ .

Such a choice of random code parameters is indeed possible: For example, by fixing  $c \geq 7$  and taking  $\mu = \frac{c^{-2}}{100}$ ,  $\delta = (\mu^c - \varepsilon_0)$ ,  $d = \frac{8\mu c^2}{(\mu^c - \delta)^2}$  we get:

$$\delta - 2\frac{\mu c}{d} = \mu^c - \varepsilon_0 - 2\frac{\mu c}{\frac{8\mu c^2}{(\mu^c - \delta)^2}} = \mu^c - \varepsilon_0 - \frac{\varepsilon_0^2}{4c}$$

Which is strictly larger than zero for all  $0 < \varepsilon_0 < 2c\left(\sqrt{1 + \frac{\mu^c}{c}} - 1\right)$ .

We now use this to prove Theorem 13:

**Proof of Theorem 13.** Fix odd integer  $c \geq 7$  and constants  $\mu, \varepsilon, \delta, d > 0$  satisfying equation (90) and  $\delta \geq \frac{\mu c}{d}$ . See Remark 37 for a specific choice of such constants. Let  $V$  be a random LDPC code of dimension  $n$  corresponding to this choice of constants. Assume that  $n$  is large enough to satisfy Lemma 36. Assume by contradiction that  $r_0 \leq n(\delta - \frac{2\mu c}{d}) - 1$ , and the Collaborative Discovery process lasts until round  $n$ . Set  $t = \lfloor \delta n \rfloor - 1$ .

At the end of round  $t$ , the coordinates  $I_t \subseteq [n]$  are disclosed. Denote by  $n_-$  the number of times a uniform distribution was predicted by the expert. Using Lemma 29, the total retention accumulated at the end of round  $t$  is equal to:

$$r_t = r_0 - n_- + (t - n_-) \quad (91)$$

$r_t \geq 0$ , and therefore  $n_- \leq \frac{t+r_0}{2}$ . Using Lemma 29 again, we obtain that  $n_- = \dim(V_{|I_t})$ . In addition, the dimensions of a vector space and its dual sum up to  $t$ , hence  $\dim((V_{|I_t})^\perp) = (t - n_-) \geq \frac{t-r_0}{2}$ . This gives us a lower bound for  $\dim((V_{|I_t})^\perp)$ .

$(V_{|I_t})^\perp$  consists of vectors  $v \in V^\perp$  such that  $\text{support}(v) \subseteq I_t$ . The conditions of Lemma 36 are satisfied by our choice of constants, and we can apply it to obtain that  $V^\perp$  of the random code we picked is  $(\delta n, \mu)$ -local with high probability, and therefore every  $v \in V^\perp$  that is a

sum of at least  $\frac{c}{d}\mu n$  dual basis vectors has  $d(v) \geq \delta n$ . For  $t < \delta n$ , all the vectors of  $(V_{I_t})^\perp$  are a sum of  $\frac{c}{d}\mu n$  basis vectors at most, hence  $\dim(V_{I_t}^\perp) \leq \frac{c}{d}\mu n$ , implying an upper bound for  $\dim((V_{I_t})^\perp)$ .

Combining the bounds we obtain:

$$\frac{t - r_0}{2} \leq \dim((V_{I_t})^\perp) < \frac{c}{d}\mu n \quad (92)$$

For  $t = \lfloor \delta n \rfloor - 1$  and  $r_0 \leq n(\delta - \frac{2\mu c}{d}) - 1$  we have:

$$\frac{t - r_0}{2} \geq \frac{(\delta n - 1) - (n(\delta - \frac{2\mu c}{d}) - 1)}{2} = \frac{c}{d}\mu n \quad (93)$$

Leading to a contradiction, since the lower bound in equation (92) must be greater than the upper bound. From this we get  $r_0 > n(\delta - \frac{2\mu c}{d})$ , and therefore  $r_0 = \Omega(n) = \Omega(\dim(V))$ . ◀

---

## References

- 1 Sanjeev Arora and Boaz Barak. *Computational complexity: a modern approach*. Cambridge University Press, 2009.
- 2 Amir Ban and Nati Linial. The dynamics of reputation systems. In *Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge*, pages 91–100. ACM, 2011.
- 3 Ayelet Ben-Sasson, Eli Ben-Sasson, Kayla Jacobs, and Eden Saig. Baby CROINC: An Online, Crowd-based, Expert-curated System for Monitoring Child Development. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*, PervasiveHealth '17, pages 110–119, New York, NY, USA, 2017. ACM. doi:10.1145/3154862.3154887.
- 4 Eli Ben-Sasson, Prahladh Harsha, and Sofya Raskhodnikova. Some 3CNF properties are hard to test. *SIAM Journal on Computing*, 35(1):1–21, 2005.
- 5 Manuel Blum, Michael Luby, and Ronitt Rubinfeld. Self-Testing/Correcting with Applications to Numerical Problems. In *Proceedings of the 22nd Annual ACM Symposium on Theory of Computing, May 13-17, 1990, Baltimore, Maryland, USA*, pages 73–83, 1990. doi:10.1145/100216.100225.
- 6 Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.
- 7 Kam Tong Chan, Irwin King, and Man-Ching Yuen. Mathematical modeling of social games. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, volume 4, pages 1205–1210. IEEE, 2009.
- 8 Alexander Philip Dawid and Monica Musio. Theory and applications of proper scoring rules. *Metron*, 72(2):169–183, 2014.
- 9 Gideon Dror, Dan Pelleg, Oleg Rokhlenko, and Idan Szpektor. Churn prediction in new users of Yahoo! answers. In *Proceedings of the 21st International Conference on World Wide Web*, pages 829–834. ACM, 2012.
- 10 Robert Gallager. Low-density parity-check codes. *IRE Transactions on information theory*, 8(1):21–28, 1962.
- 11 Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.
- 12 Michael H. Goldhaber. The attention economy and the Net. *First Monday*, 2(4), 1997. doi:10.5210/fm.v2i4.519.
- 13 Oded Goldreich, Shafi Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *Journal of the ACM (JACM)*, 45(4):653–750, 1998.

- 14 Gradient Theorem. Gradient Theorem — Wikipedia, The Free Encyclopedia, 2017. [Online; accessed 08-September-2017]. URL: [https://en.wikipedia.org/w/index.php?title=Gradient\\_theorem&oldid=781791224](https://en.wikipedia.org/w/index.php?title=Gradient_theorem&oldid=781791224).
- 15 Arlo D Hendrickson and Robert J Buehler. Proper scores for probability forecasters. *The Annals of Mathematical Statistics*, pages 1916–1921, 1971.
- 16 Richard A Lanham. *The economics of attention: Style and substance in the age of information*. University of Chicago Press, 2006.
- 17 John McCarthy. Measures of the value of information. *Proceedings of the National Academy of Sciences*, 42(9):654–655, 1956.
- 18 George A Miller. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.
- 19 Yehuda Pinchover and Jacob Rubinstein. *An introduction to partial differential equations*. Cambridge university press, 2005.
- 20 Paul Resnick, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman. Reputation systems. *Communications of the ACM*, 43(12):45–48, 2000.
- 21 Paul Resnick and Richard Zeckhauser. Trust among strangers in Internet transactions: Empirical analysis of eBay’s reputation system. In *The Economics of the Internet and E-commerce*, pages 127–157. Emerald Group Publishing Limited, 2002.
- 22 Ralph Tyrell Rockafellar. *Convex analysis*. Princeton university press, 2015.
- 23 Leonard J Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.
- 24 Endel Tulving and Fergus IM Craik. *The Oxford handbook of memory*. Oxford: Oxford University Press, 2000.
- 25 Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.
- 26 Chih-Ping Wei and I-Tang Chiu. Turning telecommunications call details to churn prediction: a data mining approach. *Expert systems with applications*, 23(2):103–112, 2002.

## A Retentive Scoring Appendices

► **Claim 21.** Let  $D \subseteq \mathbb{R}^n$  such that  $\mathbf{x}, \mathbf{y} \in D$ . For every analytic function  $u : D \times D \rightarrow \mathbb{R}$  satisfying the equation

$$u(\mathbf{x}, \mathbf{y}) - \sum_{i=1}^n (y_i - x_i) \frac{\partial u(\mathbf{x}, \mathbf{y})}{\partial x_i} = 0 \quad (32)$$

there exist functions  $\alpha_1, \dots, \alpha_n : D \rightarrow \mathbb{R}$  such that:

$$u(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n \alpha_i(\mathbf{y})(y_i - x_i) \quad (33)$$

**Proof of Claim 21.**  $u(\mathbf{x}, \mathbf{y})$  is analytic in  $D$ , and therefore it has a unique representation as a convergent power series about  $(\mathbf{y}, \mathbf{y})$ :

$$u(\mathbf{x}) = \sum_{j_1, \dots, j_{2n}=0}^{\infty} c_{j_1, \dots, j_{2n}} \prod_{k=1}^n (y_k - x_k)^{j_k} \prod_{k'=n+1}^{2n} y_{k'}^{j_{k'}} \quad (34)$$

Note that  $(y - x) \frac{\partial (y-x)^a}{\partial x} = -a(y - x)^a$  for all  $a \in \mathbb{R}$ , and therefore:

$$\sum_{i=1}^n (y_i - x_i) \frac{\partial}{\partial x_i} \prod_{k=1}^n (y_k - x_k)^{j_k} = - \sum_{i=1}^n j_i \prod_{k=1}^n (y_k - x_k)^{j_k} \quad (35)$$

Using the above, we obtain for (32):

$$0 = u + \sum_{i=1}^n (y_i - x_i) \frac{\partial u}{\partial x_i} \quad (96)$$

[Use (94) to represent the rightmost term as a power series]

$$= u + \sum_{i=1}^n (y_i - x_i) \frac{\partial}{\partial x_i} \left( \sum_{j_1, \dots, j_{2n}=0}^{\infty} c_{j_1, \dots, j_{2n}} \prod_{k=1}^n (y_k - x_k)^{j_k} \prod_{k'=n+1}^{2n} y_{k'}^{j_{k'}} \right) \quad (97)$$

[Derivative operator does not affect the factors that don't depend on  $x$ ]

$$= u + \sum_{j_1, \dots, j_{2n}=0}^{\infty} c_{j_1, \dots, j_{2n}} \prod_{k'=n+1}^{2n} y_{k'}^{j_{k'}} \left( \sum_{i=1}^n (y_i - x_i) \frac{\partial}{\partial x_i} \prod_{k=1}^n (y_k - x_k)^{j_k} \right) \quad (98)$$

[Apply the derivative using (95)]

$$= u + \sum_{j_1, \dots, j_{2n}=0}^{\infty} c_{j_1, \dots, j_{2n}} \prod_{k'=n+1}^{2n} y_{k'}^{j_{k'}} \left( - \sum_{i=1}^n j_i \right) \prod_{k=1}^n (y_k - x_k)^{j_k} \quad (99)$$

[Use (94) to represent the leftmost term as a power series]

$$= \sum_{j_1, \dots, j_{2n}=0}^{\infty} c_{j_1, \dots, j_{2n}} \left( 1 - \sum_{i=1}^n j_i \right) \prod_{k'=n+1}^{2n} y_{k'}^{j_{k'}} \prod_{k=1}^n (y_k - x_k)^{j_k} \quad (100)$$

If a convergent power series is equal to zero, then all its coefficients must be equal to zero as well. From (100) we obtain:

$$\forall j_1, \dots, j_n \in \mathbb{N} : c_{j_1, \dots, j_n} \left( 1 - \sum_{i=1}^n j_i \right) = 0 \quad (101)$$

Therefore  $c_{j_1, \dots, j_n} = 0$  when  $\sum_{i=1}^n j_i \neq 1$ , and analytic solutions for (32) can only contain linear coefficients of  $(y_i - x_i)$  in their series expansion. Let  $k \in [n]$ . when  $j_k = 1$  we denote  $c_{j_1, \dots, j_{2n}} \equiv c_{k, j_{n+1}, \dots, j_{2n}}$ . Plug back into the series representation (94) to obtain:

$$u(\mathbf{x}) = \sum_{i=1}^n \left( \sum_{j_{n+1}, \dots, j_{2n}=0}^{\infty} c_{i, j_{n+1}, \dots, j_{2n}} \prod_{k'=n+1}^{2n} y_{k'}^{j_{k'}} \right) (y_i - x_i) \quad (102)$$

Denoting  $\alpha_i(\mathbf{y}) \equiv \left( \sum_{j_{n+1}, \dots, j_{2n}=0}^{\infty} c_{i, j_{n+1}, \dots, j_{2n}} \prod_{k'=n+1}^{2n} y_{k'}^{j_{k'}} \right)$  leads to the linear representation of  $u$  in (33). ◀

## B Binary Attributes Appendices

### B.1 The Binary Attributes Model

► **Claim 26.** Let  $I \subseteq [n]$ . For every vector  $u_I \in U_I$ :

$$\Pr(\mathcal{U}_I = u_I) = 2^{-\dim(U_I)} \quad (79)$$

**Proof of Claim 26.** Without loss of generality assume that  $I = \{1, \dots, |I|\}$ , and choose a basis  $U = \text{span}\{\bar{u}_1, \dots, \bar{u}_k\}$  which is diagonalized. Each vector in  $U$  can be represented as linear combination of basis elements. By definition, only only the first  $\dim(U_I)$  diagonalized

basis vectors have support in  $I$ , and therefore every vector in  $U_I$  can be written as a linear combination of the view of the first  $\dim U_I$  basis vectors of  $U$ :

$$\forall u_I \in U_I, \exists \alpha_1, \dots, \alpha_{\dim(U_I)} : u_I = \sum_{i=1}^{\dim(U_I)} \alpha_i (\bar{u}_i)_{\downarrow I} \quad (103)$$

Picking  $u$  at random is equivalent to choosing each  $\alpha_i$  uniformly, or equivalently, picking  $(\alpha_1, \dots, \alpha_{\dim(U_I)}) \sim \text{Uniform}(\{0, 1\}^{\dim(U_I)})$ . From this correspondence it follows that  $\Pr(u_I) = \Pr(\alpha_1, \dots, \alpha_{\dim(U_I)}) = 2^{-\dim(U_I)}$ .  $\blacktriangleleft$

► **Claim 27.** Let  $I \subseteq [n]$  and  $m \in [n] \setminus I$ , and assume a vector  $u \in \mathbb{F}_2^n$  has been picked uniformly at random from a vector space  $U$ .  $\Pr(u_m \mid u_I)$  is a singleton distribution if and only if  $e_m \in U^\perp_{\downarrow [n] \setminus I}$ .

**Proof of Claim 27.** When  $e_m \in U^\perp_{\downarrow [n] \setminus I}$  there exists a vector  $v \in U^\perp$  and  $I' \subseteq I$  such that  $\text{support}(v) = \{m\} \cup I'$ .  $v$  is a dual-space vector, and therefore  $\sum_{i \in I'} u_i + u_m = 0$ . The value  $u_m \in \{0, 1\}$  is completely determined by the values of  $u_{I'}$ , and therefore  $\Pr(u_m \mid u_I)$  is a singleton distribution.

Conversely, observe that restricting a vector to a subset of coordinates  $I \subseteq [n]$  can be viewed as a linear projection operation  $P_I \equiv \sum_{i \in I} e_i e_i^T$ . Let  $v \in U$  be a vector for which  $v_I = u_I$ . The set of vectors  $u' \in U$  for which  $u'_I = u_I$  is an affine subspace  $U'$  of  $U$ :

$$U' = v + V' = \{v + v' \mid v' \in U, P_I v' = 0\} \quad (104)$$

Note that  $V'$  is a linear subspace of  $U$ , and therefore:

$$(V')^\perp = \text{span}(U^\perp \cup \{e_i \mid i \in I\}) \quad (105)$$

Using the assumption that  $\Pr(u_m \mid u_I)$  is a singleton distribution, we get that the  $m$ -th coordinate is constant in  $U'$ , and therefore  $P_{\{m\}} V' = 0$ , and  $e_m \in (V')^\perp$ . denote  $U^\perp = \text{span}\{\bar{u}_1^\perp, \dots, \bar{u}_{n-k}^\perp\}$ . Using (105) we can write  $e_m$  as a linear combination of spanning set elements:

$$e_m = \sum_{i=1}^{|I|} \alpha_i e_i + \sum_{j=1}^{n-k} \beta_j \bar{u}_j^\perp \quad (106)$$

Restricting the view to coordinates  $[n] \setminus I$ , the terms in the first sum vanish, yielding:

$$e_m = P_{[n] \setminus I} e_m = \sum_{j=1}^{n-k} \beta_j P_{[n] \setminus I} \bar{u}_j^\perp \quad (107)$$

We have shown that it's possible to write  $e_m$  as a linear combination of punctured dual space elements, hence  $e_m \in U^\perp_{\downarrow [n] \setminus I}$ .  $\blacktriangleleft$

► **Claim 28.** Let  $U$  be a linear space over  $\mathbb{F}_n^2$ , and let  $I \subseteq [n], m \in [n] \setminus I$ .  $e_m \in U^\perp_{\downarrow [n] \setminus I}$  if and only if  $\dim(U_{\downarrow I}) = \dim(U_{\downarrow I \cup \{m\}})$ .

**Proof of Claim 28.** Assume a uniform distribution over  $U$ , then  $e_m \in U^\perp_{\downarrow [n] \setminus I}$ , if and only if  $\Pr(u_m \mid u_I)$  is a singleton distribution by Claim 27.

According to the law of total probability,  $\Pr(u_m \mid u_I)$  is a singleton distribution if and only if the following marginal distributions are equal:  $\Pr(u_{I \cup \{m\}}) = \Pr(u_I)$ .

Using Claim 26 we obtain that the two probabilities are equal if and only if  $\dim(U_{\downarrow I}) = \dim(U_{\downarrow I \cup \{m\}})$ .  $\blacktriangleleft$



## B.2 Retention Complexity of the Walsh-Hadamard Code

► **Claim 32.** Let  $y^{(1)}, \dots, y^{(m)} \in (\{0, 1\}^k \setminus \{0^k\})$ .

$$\left( \sum_{i=1}^m e_{y^{(i)}} \right) \in \text{WH}^\perp \iff \sum_{i=1}^m y^{(i)} = 0 \quad (87)$$

**Proof of Claim 32.** By definition,  $(\sum_{i=1}^m e_{y^{(i)}}) \in \text{WH}^\perp$  if and only if  $(\sum_{i=1}^m e_{y^{(i)}}) \cdot u = 0$  for all  $u \in \text{WH}$ . For an arbitrary  $u$ , let  $w \in \{0, 1\}^k$  such that  $u = \text{WH}(w)$ . Plug into the definition of WH and obtain:

$$\begin{aligned} \left( \sum_{i=1}^m e_{y^{(i)}} \right) \cdot u &= \sum_{i=1}^m u_{y^{(i)}} \\ &= \sum_{i=1}^m w \cdot y^{(i)} \\ &= w \cdot \left( \sum_{i=1}^m y^{(i)} \right) \end{aligned}$$

Observe that the inner product is equal to zero for all  $u \in \text{WH}$  if and only if  $w \cdot (\sum_{i=1}^m y^{(i)}) = 0$  for all  $w \in \{0, 1\}^k$ . This happens if and only if  $(\sum_{i=1}^m y^{(i)}) = 0$ , proving our claim. ◀

► **Claim 33.**

$$d(\text{WH}^\perp) = 3 \quad (88)$$

**Proof of Claim 33.** By Claim 32, the vectors corresponding to the support of each constraint in  $\text{WH}^\perp$  must have their XORs equal to zero.

$0^k \notin (\{0, 1\}^k \setminus \{0^k\})$ , and therefore there are no constraints of size 1, and we have  $d(\text{WH}^\perp) > 1$ . Similarly, for all  $x, y \in (\{0, 1\}^k \setminus \{0^k\})$  such that  $x \neq y$  we get  $x + y \neq 0$ , and therefore there are no constraints of size 2, and  $d(\text{WH}^\perp) > 2$ .

Taking  $x \neq y$  and  $z = x + y$  gives 3 coordinates with corresponding vectors that sum up to zero, and therefore  $d(\text{WH}^\perp) \leq 3$  according to Claim 32. Combining the conclusions we obtain  $d(\text{WH}^\perp) = 3$ . ◀