Semantic Compression with Information Lattice Learning

Haizi Yu and Lav R. Varshney
University of Illinois Urbana-Champaign

***Abstract*—Data-driven artificial intelligence (AI) techniques are becoming prominent for learning in support of data compression, but are focused on standard problems such as text compression. To instead address the emerging problem of semantic compression, we argue that the lattice theory of information is particularly expressive and mathematically precise in capturing notions of abstraction as a form of lossy semantic compression. As such, we demonstrate that a novel AI technique called information lattice learning, originally developed for knowledge discovery and creativity, is powerful for learning to compress in a semantically-meaningful way. The lattice structure further implies the optimality of group codes and the successive refinement property for progressive transmission.**

I. Introduction

There has been growing interest in using large language models (LLMs) and similar large artificial intelligence (AI) models in data compression [1]. Such work has focused on standard compression problems, but there is also growing interest in semantic communication and information representation [2] especially motivated by 6G wireless communication systems [3]. In this work, we take up the challenge of learning to compress in a semantically meaningful way and particularly argue that *abstraction* is a very natural approach to lossy semantic compression. In fact, abstractions are core to language and cognition of meaning [4], [5], and communicative needs are often why abstractions are learned in the first place [6]. Rather than LLMs, we will see that an alternative approach to large-scale AI called *information lattice learning* [7] is well-matched to semantic compression.

A recent textbook on the linguistics of meaning notes that the unit of meaning in semantics is the *proposition*, but propositions are "probably the most recalcitrant constructs to define" [8, p. 6]. Propositions can be thought of as descriptions of situations or contents of beliefs. As indicated in [8], (1) and (2) express the same proposition:

$$\text{The dog has eaten the roti.} \tag{1}$$

$$\text{The roti has been eaten by the dog.} \tag{2}$$

even though a sentence in active voice and its passive voice equivalent do not contain identical information (since voice impacting semantic role of words is also a piece of information contained in the sentence). Further, (3) and its French translation (4):

$$\text{I am happy.} \tag{3}$$

$$\text{Je suis content.} \tag{4}$$

express the same proposition if they are accurate translations.

In a fairly obscure work entitled "The Lattice Theory of Information," Claude Shannon aimed to study the nature of information rather than just its amount (as in his famous 1948 paper [9]), and arrived at the same concept [10]. As he said, "Suppose a source is producing, say, English text. This may be translated or encoded into many other forms (e.g. Morse code) in such a way that it is possible to decode and recover the original. For most purposes of communication, any of these forms is equally good and may be considered to contain the same information.... Each coded version of the original process may be called a *translation* of the original language. These translations may be viewed as different ways of describing the same information, in about the same way that a vector may be described by its components in various coordinate systems. The information itself may be regarded as the equivalence class of all translations or ways of describing the same information."

He basically defined an *information element* as the equivalence class of random variables that induce the same σ-algebra. For further explication, see [11], [12]. The notion of an information element is more abstract than that of a random variable: an information element can be realized by different random variables. With this mathematization, numerous algebraic and topological properties follow. Further, the notion of a partial order among information elements and the resultant *information lattices* arise naturally from the definition of information element, providing a hierarchical depiction of information elements at different abstraction levels. Notably, the fact every σ-algebra of a countable sample space can be uniquely determined by its generating sample-space-partition implies that every information lattice is isomorphic to its underlying partition lattice. Since a partition of a sample space is essentially an equivalence relation, the abstraction hierarchy depicted by an information lattice can naturally yield semantic abstractions and hierarchies (induced by various equivalence relations, e.g., by a subgroup lattice)—both in a human-interpretable manner.

In this paper, we argue that the concept of information elements in the lattice theory of information is a natural mathematical formulation of propositions in semantics, and accordingly, information lattices are natural for foundational work in the emerging area of *semantic communication*. Although going back in some ways to the work of Bar-Hillel and Carnap [13], there has been renewed interest in semantic communication and its formal foundations in recent years [2],

[14].

While a simple extension of information theory to semantic communication based on synonym mappings has recently been proposed [15], the information lattice approach enables a significant rethinking of lossless and lossy information representation when taking semantics into account.

Shannon's original conception of information lattices assumes the probability measure is given. Yet, learning statistics from data enables matching of compression schemes to complicated source statistics for improved performance in many modern compression applications [16]. As such, we further argue that *information lattice learning*—as we have developed in a sequence of recent papers for knowledge discovery and creativity [7], [17]–[23]—is well-suited to learn information lattices (essentially, hierarchical semantic abstractions) for given sources and then to compress in semantically-meaningful ways with human-understandable and mathematically precise fidelity criteria.

The implications of information lattice learning for semantic compression can be significant. Given the group-theoretic foundations of information lattices, one can remember the exponential rate savings that are possible in lossless and lossy data compression under permutation group invariance (corresponding to the semantics of scientific data and other similar sources) [24]. (Group-theoretic ideas have also become prominent in AI to reduce societal bias and sample complexity, e.g. [25].) Further, one can remember that group lattice structures enable perfect progressive transmission, with no rate loss in the multiple descriptions and successive refinement problems [26]. Formalizing distortion using a lattice-based distance measure for partitions, we show the same kind of results for more general semantic compression with information lattice learning.

The remainder of the paper is organized as follows. Sec. II gives a discursive review/intuition of Shannon's information lattices and our information lattice learning, so as to demonstrate why the information lattice framework is natural for semantic compression; it also presents natural fidelity criteria from within this framework. Sec. III shows examples to help clarify lossy semantic compression as an abstraction process. Sec. IV considers the successive refinement problem for semantic compression in the information lattice framework, showing there is no rate loss using group codes due to the lattice structure. Sec. V concludes the paper.

II. Information Lattices and Information Lattice Learning

Starting from a standard setup in probability theory, let $(\Omega, \mathcal{F}, P)$ be a probability space, where Ω is called a sample space consisting of all possible outcomes and $\mathcal{F}$ is a σ-algebra (on Ω) consisting of all measurable events. A random variable is a measurable function $\mathbf{X} : (\Omega, \mathcal{F}) \to (X, \mathcal{X})$.

Intuitively, one can view $(\Omega, \mathcal{F})$ as some topic space, where the σ-algebra $\mathcal{F}$ defines the full information about this topic. One can then view a random variable $\mathbf{X}$ as a language (e.g., English) and attempt to describe the topic in that language. In particular, $\mathbf{X}$'s codomain X can be thought of as the full vocabulary of the language (with $\mathbf{X}(\Omega)$ being the part of the vocabulary that is related to the topic) and its σ-algebra $\mathcal{X}$ as everything describable by that language.

Related to a particular topic, a language may not be expressive enough in describing the topic. We see scenarios where we have "lack of words" when attempting to clearly and precisely describe a topic, e.g., attempting to describe music imagined in the mind or a bodily feeling to a doctor. Moreover, not all languages are equally expressive in describing the topic. A novel originally written in one language may lose information when translated to another. Similarly, the Sapir–Whorf hypothesis states that "individuals' languages determine or shape their perceptions of the world" [27].

The above intuition can be naturally formalized via induced σ-algebras. Given a topic $(\Omega, \mathcal{F})$, we define the *descriptive power* of a language $\mathbf{X}$ with respect to the topic to be $\sigma(\mathbf{X})$, i.e., the σ-algebra on Ω induced by $\mathbf{X}$. Since a random variable is a measurable function, $\sigma(\mathbf{X}) \subseteq \mathcal{F}$. This includes two cases:

- $\sigma(\mathbf{X}) = \mathcal{F}$: the language $\mathbf{X}$ describes the topic losslessly;
- $\sigma(\mathbf{X}) \subset \mathcal{F}$: lossy description.

Given two languages as two distinct random variables $\mathbf{X} : (\Omega, \mathcal{F}) \to (X, \mathcal{X})$ and $\mathbf{Y} : (\Omega, \mathcal{F}) \to (Y, \mathcal{Y})$, we can check their respective descriptive power $\sigma(\mathbf{X})$ and $\sigma(\mathbf{Y})$ for the given topic $(\Omega, \mathcal{F})$. To name a few possibilities,

(a) $\sigma(\mathbf{X}) = \sigma(\mathbf{Y})$: informationally equivalent;
(b) $\sigma(\mathbf{X}) \subseteq \sigma(\mathbf{Y}) \subseteq \mathcal{F}$: $\mathbf{X}$ is more informationally lossy compared to $\mathbf{Y}$. More precisely, $\mathbf{Y}$ is informationally lossy when describing the whole topic $(\Omega, \mathcal{F})$ but can describe all $\mathbf{X}$ can describe.
(c) $\sigma(\mathbf{X}) \not\subseteq \sigma(\mathbf{Y})$ and $\sigma(\mathbf{Y}) \not\subseteq \sigma(\mathbf{X})$: non-comparable.

The above suggests that the nature of information conveyed by a random variable is not really the random variable *per se*, but its induced σ-algebra. This leads to Shannon's introduced *information element* capturing such *nature of information*. Hence, one does not need to rely on a particular language or random variable, but can instead, directly use an information element to faithfully refer to a piece of information.

A. Information Element and Information Lattice

Formally, an *information element* is an equivalence class of random variables (on a common sample space), where the equivalence relation ($\sim$) is defined as follows: two random variables $\mathbf{X} \sim \mathbf{Y}$ if and only if $\sigma(\mathbf{X}) = \sigma(\mathbf{Y})$.

The relationship in (b) and (c) above naturally leads to a partial order among information elements (based on the $\subseteq$ relation among σ-algebras), and this partial order can be further proven to form a lattice of information elements, or *information lattice* in short.

Unlike in, but equivalent to Shannon's original definition, the above definition of information element and lattice is stated more generally using just σ-algebras, which in particular does not require probability or entropy to be predefined. This suggests a breakdown of an information element into two parts: σ-algebra and probability measure, and accordingly, a breakdown

of an information lattice into a σ-algebra lattice and probability measures, where the σ-algebra lattice is isomorphic to the information lattice. Note that in cases where the sample space Ω is countable, every σ-algebra on Ω bijectively corresponds to its generating partition of Ω, so, an information lattice is also isomorphic to its underlying partition lattice (equipped with the usual "coarser/finer than" partial order).

B. Information Lattice Learning (ILL)

The separation of probability measures from a σ-algebra lattice, or equivalently a partition lattice in the countable case, brings learning into the information lattice, yielding the general framework called *information lattice learning*. In this framework, learning can happen in two directions, following the forward and backward direction of the partial order. From the top of a lattice (corresponding to the original σ-algebra $\mathcal{F}$), one can *project* a probability measure, or any signal, down to the lower parts of the lattice to learn coarser-grained summarizing signals (also known as *rules* in information lattice learning)—a lossy compression process. Conversely, one can *lift* coarser-grained summarizing signals (or rules) up to learn a finer-grained realization signal, e.g., learning a probability distribution that satisfies the rules—a reconstruction or decompression process. The transparency of information lattices and lattice learning allows semantic compression/decompression to be done in a human-interpretable manner.

C. Semantic Fidelity Criteria

As detailed in [22], Shannon originally defined an entropy-based distance measure within information lattices (with probabilities already attached): for any two information elements x and y, the distance between them $\rho(x, y)$ is defined to be the sum of two conditional entropies: $\rho(x, y) := H(x|y)+H(y|x)$. Also note that many information theory textbooks, e.g. [28], [29], give Venn diagrams called i-diagrams that show conditional entropy as a symmetric difference.

Here we consider more generic settings of ILL where statistics may not be attached to a σ-algebra lattice or a partition lattice yet. In particular, we consider a lattice-based distance measure for partitions [30], which is suitable for establishing semantic fidelity criteria. More specifically, given a probability space $(\Omega, \mathcal{F}, P)$, consider the lattice of all partitions of the sample space Ω. One can define a partition distance as follows: for any partitions P, Q of Ω,

$$\delta_L(P, Q) := |P \wedge Q| - |P \vee Q|,$$

where $P \wedge Q$ is the coarsest common refinement of P, Q (also known as the *meet*) and $P \vee Q$ is the finest common coarsening of P, Q (also known as the *join*).

III. Examples

We provide two examples to explicate the notion of abstraction as a form of lossy semantic compression. The reader is encouraged to think through the relevance of the partition distance as an appropriate fidelity criterion in these examples.

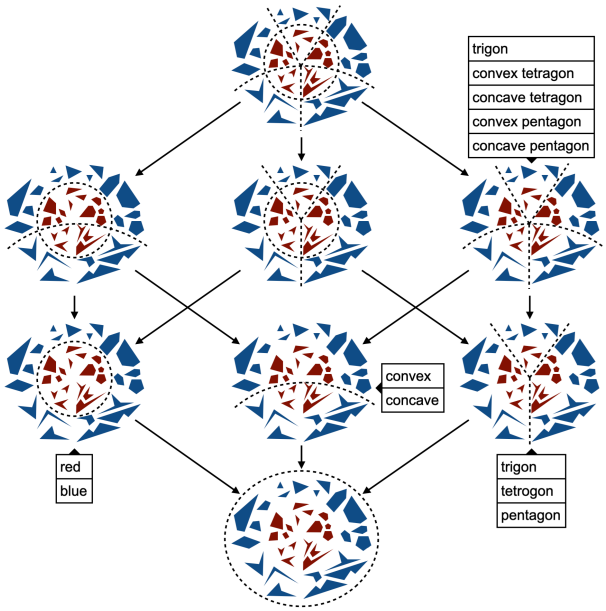

Fig. 1. A lattice of partitions of a set of polygons.

First, consider the representation of shapes, which arises in 6G wireless applications such as virtual reality and augmented reality [31]. In Fig. 1, we see data that corresponds to shapes of various kinds, and the partition lattice that is constructed according to the concepts of color, convexity, and number of sides. Notice that going down the lattice from, say, convex pentagons, can lead to a coarsened lossy representation into either pentagons or into convex shapes. Such abstraction operates directly in the semantic space of meaning.

A more detailed example of using ILL in lossy semantic compression is to learn hierarchical semantic abstractions of raw music information, e.g., encoded in sheet music. We explicitly constructed partition lattices from symmetries like isometries (rigid body transformations such as translation and rotation) and then trained the information lattice for music on the basis of a fairly small amount of data, just 370 chorales by Johann Sebastian Bach. This procedure recovers a large fraction of the laws of music theory in the same human-interpretable form as an undergraduate textbook and further discovers new principles of interest to music theorists [7], [20].

To illustrate lossy semantic compression for this example, consider the slightly modified excerpt from Mozart's Piano Sonata No. 16 (K 545) at the top of Fig. 2. One can inject the excerpt at the top of an information lattice trained for music; then projecting it down along the lattice yields hierarchical lossy compressions corresponding to several human-interpretable music concepts at different levels of semantic abstractions. From top down in Fig. 2, sampling is a fairly direct lossy semantic compression which directly drops unwanted parts and is commonly seen in music production and mixing (e.g., among DJs). One can further go down through different compression paths to yield different types of music semantic abstractions that are not directly comparable to each

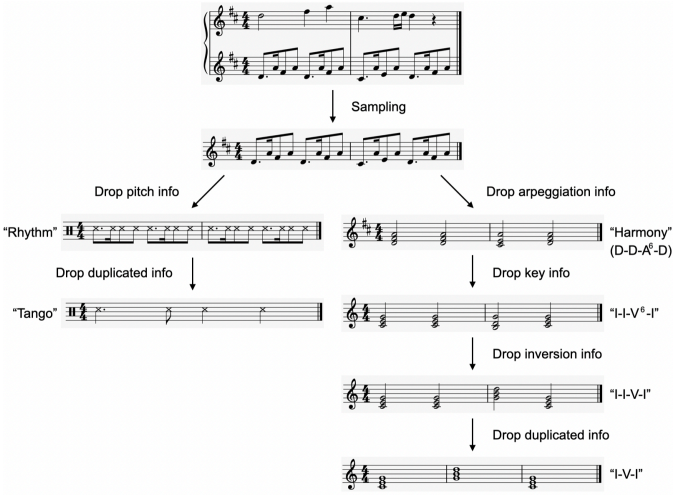

Fig. 2. Lossy semantic music compression along an information lattice (the complete lattice is not shown for brevity).

other. For example, dropping pitch information allows lossy compression of music to contain just its rhythm information, while dropping arpeggiation information allows a different type of lossy compression of music to contain just its harmony information. Along each compression path in an information lattice, one can continuously drop information to yield more and more lossy compressions, corresponding to deeper and deeper levels of music concepts, such as harmonic progression related to a particular key, roman numerals (with and without figures), harmonic functions, and so on.

IV. Group Codes and Progressive Transmission

In addition to single descriptions in semantic compression, there may also be interest in progressive transmission [32], e.g. in 6G wireless systems, so each received packet provides semantic insight and further packets build up towards more and more semantic information. There is some nascent work on semantic multiresolution representation [33], but this topic largely remains unexplored. In particular, we consider the successive refinement framework in rate-distortion theory [34], [35]. Note that the fact there is no rate loss in successive refinement for lossless representation is direct, cf. [36].

Besides the motivation from wireless networks, the successive refinement setting also models the incorporation of new information into already learned representations, as in developmental learning, lifelong learning, or indeed in any kind of learning process that progresses through identifiable successive steps [37] such as epoch-based training and ILL's rule learning itself. As such, it is of interest to know whether semantic information can be decomposed into chunks without needing extra rate.

As noted in Sec. II, any information lattice is isomorphic to its underlying σ-algebra lattice, or partition lattice in the countable case. Moreover, any partition lattice is isomorphic to a subgroup lattice in group theory [11]. In a subgroup lattice, the join of two subgroups is the subgroup generated by their union, and the meet of two subgroups is their intersection. Cayley's Theorem states that every group is isomorphic to a permutation group, and so in a sense, permutation groups play a special role in group theory. For brevity and ease of explanation, here we restrict our attention to permutation groups and ask whether lossy semantic compression codes for the subgroup lattice of a permutation group (see Fig. 3 for its depiction as a partition lattice for a source set of four items $\Omega = \{\alpha, \beta, \gamma, \zeta\}$) have the successive refinement property in rate-distortion theory.

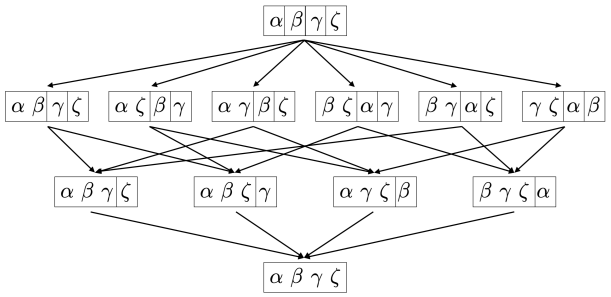

Fig. 3. A partition lattice for a four-item set corresponding to a subgroup lattice of a permutation group.

These codes will be permutation source codes [38], [39]: the basic idea of these group source codes is to represent partial orders for source sequences.

A. Distortion Measure

Let us first introduce notation and concepts specifically for orders and permutations, based on [26]. Then we will connect to more general notions for ILL, such as the distance-based notion of distortion within partition lattices.

Recall that a binary relation $\leq$ on a set Ω is a *partial order* if it satisfies the reflexive ($x \leq x$ for all $x \in \Omega$), transitive ($x \leq x'$ and $x' \leq x''$ implies $x \leq x''$ for all $x, x', x'' \in \Omega$), and antisymmetric ($x \leq x'$ and $x' \leq x$ implies $x = x'$ for all $x, x' \in \Omega$) properties. A partial order satisfying comparability (for any x, x' in Ω, either $x \leq x'$ or $x' \leq x$) is a *total order*.

As an example for $n = 4$, let $\Omega = \{\alpha, \beta, \gamma, \zeta\}$ equipped with the total order $k = \{\alpha \leq \beta \leq \gamma \leq \zeta\}$. Let O be the set of all partial orders on Ω that is consistent with k. A partial order j is consistent with k if all comparable pairs in j exist in k. For any partial order $j \in O$, we encode j as a binary *comparability vector* of length $n-1$: $\tilde{j} = (\mathbb{1}_{[\alpha \leq \beta]}, \mathbb{1}_{[\beta \leq \gamma]}, \mathbb{1}_{[\gamma \leq \zeta]})$, representing knowledge of comparability. This allows us to define a distortion measure δ_n between any partial order j on Ω and the total order k as follows:

$$\delta_n(j) = \begin{cases} \frac{1}{n-1} d_H(\tilde{j}, \tilde{k}), & j \in O \\ \infty, & j \notin O. \end{cases} \tag{5}$$

where $d_H(\cdot, \cdot)$ is Hamming distance.

TABLE I
PERMUTATION CODES AND COMPARABILITY RELATIONSHIPS DETERMINED BY CORRESPONDING PARTIAL ORDERS

$[1,1,1,1]$	$\{a,b,c\}$	U
$[2,1,1]$	$\{b,c\}$	F
$[1,2,1]$	$\{a,c\}$	E
$[1,1,2]$	$\{a,b\}$	D
$[2,2]$	$\{b\}$	B
$[3,1]$	$\{c\}$	C
$[1,3]$	$\{a\}$	A
$[4]$	$\emptyset$	O

For example, the distortion between a partial order in the following

$$j = \{\}$$
$$j' = \{\alpha \leq \beta\}$$
$$j'' = \{\alpha \leq \beta \leq \gamma,\, \beta \leq \zeta\}$$
$$j''' = \{\alpha \leq \beta \leq \zeta \leq \gamma\}$$

and the total order k satisfies $\delta_4(j) = 1$, $\delta_4(j') = 2/3$, $\delta_4(j'') = 1/3$, $\delta_4(j''') = \infty$; note $\delta_4(k) = 0$. Erasure of comparability knowledge incurs finite distortion, but error incurs infinite distortion. Without loss of optimality, a source code never uses inconsistent reproductions, since the order with no defined comparability relations is consistent with all total orders and has maximum distortion 1.

Assuming no inconsistent reproductions, one may check that (5) is equivalent to the partition distance defined in Sec. II-C (such equivalence can be seen from the partition generated by a comparability vector, e.g., $(0,1,0) \mapsto \{\{\alpha,\beta\},\{\gamma,\zeta\}\}$). Within a connected path up the partition lattice for a permutation group, this is governed by counting the number of comparability relations that need to be established.

B. Permutation Codes

We represent partial orders for semantic compression in the setting of the permutation group, exactly what is accomplished by permutation source codes [38], [39]. If one uses a permutation code with block size n and pool sizes $(n_1, \ldots, n_K)$, then the members within each pool are incomparable, whereas the pools themselves are comparable. Due to the intimate relationship between the pool sizes for a permutation code and the number of comparability relations established, clearly permutation codes achieve the rate-distortion limit for single descriptions.

Table I gives the possible permutation codes (for the $n = 4$ case) from the power set of possible comparability relations:

$$\alpha \overset{a}{\leq} \beta \overset{b}{\leq} \gamma \overset{c}{\leq} \zeta.$$

This may also be drawn as a lattice of codes as in Fig. 4a, corresponding to the partition lattice in Fig. 4b. Nodes on the same level of the lattice of codes have the same distortion. Since reaching different lattice nodes requires different rates, source coding simply involves choosing the low-rate node on the desired distortion level. This is optimal.

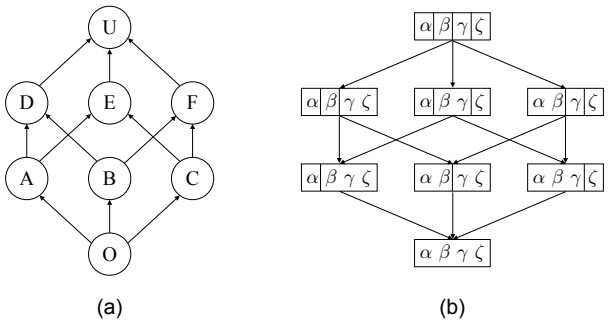

Fig. 4. Lattice of codes and its corresponding partition lattice based on comparability relations to be determined, where $\mathtt{O} = \emptyset$, $\mathtt{A} = \{a\}$, $\mathtt{B} = \{b\}$, $\mathtt{C} = \{c\}$, $\mathtt{D} = \{a,b\}$, $\mathtt{E} = \{a,c\}$, $\mathtt{F} = \{b,c\}$, and $\mathtt{U} = \{a,b,c\}$.

The lattice of codes implies that successive refinement involves choosing paths. That is, sub-permutation codes may be used to define ordering among pool elements. The sub-permutation code for the members of the first pool would be of blocklength n_1 and pool sizes $(m_1, \ldots, m_L)$. When this refinement information is used to supplement the original permutation code, the distortion is exactly equivalent to a blocklength n, pool size $(m_1, \ldots, m_L, n_2, \ldots, n_K)$ permutation code. The total rate for the two-step procedure is

$$\begin{aligned} R_{sr} &= \log \frac{n!}{\prod_{i=1}^{K} n_i!} + \log \frac{n_1!}{\prod_{j=1}^{L} m_i!} \\ &= \log \frac{n! n_1!}{\prod_{j=1}^{L} m_i! \prod_{i=1}^{K} n_i!} = \log \frac{n!}{\prod_{j=1}^{L} m_i! \prod_{i=2}^{K} n_i!}. \end{aligned} \tag{6}$$

The total rate for a single description permutation code of the same performance would be identical to (6). By repeated application of this property, there is no rate loss in successively refining a permutation code with a sub-permutation code, which achieves optimal rate-distortion.

One can extend this argument from successive refinement to multiple descriptions, following [26]. Future work aims to show successive refinability and multiple descriptions optimality of general ILL-based semantic compression using group source codes [40].

V. CONCLUSION

With strong information-theoretic and group-theoretic foundations, ILL is a novel non-neural approach to machine learning, emphasizing transparency of the model (non-blackbox) and human-interpretability of what has been learned by the model rather than just model performance on specific end tasks. In this paper, we have argued that it is a natural approach to semantic data compression and also readily implemented using group source codes applied on top of lattices learned from data. Going forward, it is of interest to develop this approach for representing semantic meaning in a variety of application areas [41], [42] and to demonstrate superior rate-distortion performance.

REFERENCES

[1] C. S. K. Valmeekam, K. Narayanan, D. Kalathil, J.-F. Chamberland, and S. Shakkottai, "LLMZip: Lossless text compression using large language models," arXiv:2306.04050 [cs.IT]., Jun. 2023.

[2] D. Gündüz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 5–41, Jan. 2023.

[3] Y. E. Sagduyu, T. Erpek, A. Yener, and S. Ulukus, "Will 6G be semantic communications? opportunities and challenges from task oriented and secure communications to integrated sensing," arXiv:2401.01531 [cs.NI]., Jan. 2024.

[4] D. Kayser, "Abstraction and natural language semantics," *Phil. Trans. R. Soc. B, Biol. Sci.*, vol. 358, no. 1435, pp. 1261–1268, Jul. 2003.

[5] F. Pulvermüller, "How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics," *Trends Cogn. Sci.*, vol. 17, no. 9, pp. 458–470, Sep. 2013.

[6] D. L. Schwartz, "The emergence of abstract representations in dyad problem solving," *J. Learning Sci.*, vol. 4, no. 3, pp. 321–354, 1995.

[7] H. Yu, J. A. Evans, and L. R. Varshney, "Information lattice learning," *J. Artif. Intell. Res.*, vol. 77, pp. 971–1019, Jul. 2023.

[8] K. M. Jaszczolt, *Semantics, Pragmatics, Philosophy: A Journey through Meaning*. Cambridge, UK: Cambridge University Press, 2023.

[9] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, July/Oct. 1948.

[10] ——, "The lattice theory of information," *Trans. IRE Prof. Group Inf. Theory*, vol. 1, no. 1, pp. 105–107, Feb. 1953.

[11] H. Li and E. K. P. Chong, "Information lattices and subgroup lattices: Isomorphisms and approximations," in *Proc. 45th Annu. Allerton Conf. Commun. Control Comput.*, Sep. 2007, pp. 1103–1110.

[12] I. Delsol, O. Rioul, J. Béguinot, V. Rabiet, and A. Souloumiac, "An information theoretic condition for perfect reconstruction," *Entropy*, vol. 26, no. 1, p. 86, Jan. 2024.

[13] Y. Bar-Hillel and R. Carnap, "Semantic information," *Br. J. Philos. Sci.*, vol. 4, no. 14, pp. 147–157, Aug. 1953.

[14] Y. Shao, Q. Cao, and D. Gündüz, "A theory of semantic communication," arXiv:2212.01485 [cs.IT]., Dec. 2022.

[15] K. Niu and P. Zhang, "A mathematical theory of semantic communication," arXiv:2401.13387 [cs.IT]., Jan. 2024.

[16] E. Ozyilkan and E. Erkip, "Distributed compression in the era of machine learning: A review of recent advances," in *Proc. 58th Annu. Conf. Inf. Sci. Syst. (CISS 2024)*, Mar. 2024.

[17] H. Yu and L. R. Varshney, "Towards deep interpretability (MUS-ROVER II): Learning hierarchical representations of tonal music," in *Proc. 6th Int. Conf. Learn. Represent. (ICLR)*, Apr. 2017.

[18] H. Yu, I. Mineyev, and L. R. Varshney, "Orbit computation for atomically generated subgroups of isometries of $\mathbb{Z}^n$," *SIAM J. Appl. Algebr. Geom.*, vol. 5, no. 3, pp. 479–505, Sep. 2021.

[19] H. Yu, H. Taube, J. A. Evans, and L. R. Varshney, "Human evaluation of interpretability: The case of AI-generated music knowledge," in *ACM CHI 2020 Workshop Artif. Intell. HCI: A Modern Approach*, Apr. 2020.

[20] H. Yu, L. R. Varshney, H. Taube, and J. A. Evans, "(Re)discovering laws of music theory using information lattice learning," *IEEE BITS Inf. Theory Mag.*, vol. 2, no. 1, pp. 58–75, Oct. 2022.

[21] H. Yu, J. A. Evans, D. Gallo, A. J. Kruse, W. M. Patterson, and L. R. Varshney, "AI-aided co-creation for wellbeing," in *Proc. 2nd Workshop Future Co-Creative Syst.*, Sep. 2021, pp. 453–456.

[22] H. Yu, I. Mineyev, and L. R. Varshney, "A group-theoretic approach to computational abstraction: Symmetry-driven hierarchical clustering," *J. Mach. Learn. Res.*, vol. 24, no. 47, pp. 1–61, 2023.

[23] H. Yu, I. Mineyev, L. R. Varshney, and J. A. Evans, "Learning from one and only one shot," arXiv:2201.08815 [cs.CV]., Jan. 2022.

[24] L. R. Varshney and V. K. Goyal, "Toward a source coding theory for sets," in *Proc. IEEE Data Compression Conf. (DCC 2006)*, Mar. 2006, pp. 13–22.

[25] S. Basu, P. Sattigeri, K. Natesan Ramamurthy, V. Chenthamarakshan, K. R. Varshney, L. R. Varshney, and P. Das, "Equi-tuning: Group equivariant fine-tuning of pretrained models," in *Proc. 37th AAAI Conf. Artif. Intell.*, Feb. 2023, pp. 6788–6796.

[26] L. R. Varshney and V. K. Goyal, "Ordered and disordered source coding," in *Proc. Inf. Theory Appl. Inaugural Workshop*, Feb. 2006.

[27] E. Sapir, *Language: An Introduction to the Study of Speech*. Harcourt, Brac, 1921.

[28] F. Reza, *An Introduction to Information Theory*. New York: McGraw-Hill, 1961.

[29] R. W. Yeung, *A First Course in Information Theory*. New York: Kluwer Academic/Plenum Publishers, 2002.

[30] G. Rossi, "Partition distances," arXiv:1106.4579 [cs.DM]., Jun. 2011.

[31] Y. Zhang, K. Ding, N. Li, H. Wang, X. Huang, and C.-C. J. Kuo, "Perceptually weighted rate distortion optimization for video-based point cloud compression," *IEEE Trans. Image Process.*, vol. 32, pp. 5933–5947, Oct. 2023.

[32] V. K. Goyal, "Multiple description coding: Compression meets the network," *IEEE Signal Process. Mag.*, vol. 18, no. 5, pp. 74–93, Sep. 2001.

[33] M. Mortaheb, M. A. A. Khojastepour, S. T. Chakradhar, and S. Ulukus, "Semantic multi-resolution communications," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM 2023)*, Dec. 2023.

[34] V. N. Koshelev, "Hierarchical coding of discrete sources," *Probl. Inf. Transm.*, vol. 16, no. 3, pp. 31–49, 1980.

[35] W. H. R. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Trans. Inf. Theory*, vol. 37, no. 2, pp. 269–275, Mar. 1991.

[36] L. R. Varshney, J. Kusuma, and V. K. Goyal, "Malleable coding for updatable cloud caching," *IEEE Trans. Commun.*, vol. 64, no. 12, pp. 4946–4955, Dec. 2016.

[37] H. Charvin, N. C. Volpi, and D. Polani, "Exact and soft successive refinement of the information bottleneck," *Entropy*, vol. 25, no. 9, p. 1355, 2023.

[38] T. Berger, F. Jelinek, and J. K. Wolf, "Permutation codes for sources," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 1, pp. 160–169, Jan. 1972.

[39] V. K. Goyal, S. A. Savari, and W. Wang, "On optimal permutation codes," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 2961–2971, Nov. 2001.

[40] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices, and Groups*. New York: Springer-Verlag, 1998.

[41] P. Gärdenfors, *Geometry of Meaning: Semantics Based on Conceptual Spaces*. Cambridge, MA, USA: MIT Press, 2014.

[42] A. O. Constantinescu, J. X. O'Reilly, and T. E. J. Behrens, "Organizing conceptual knowledge in humans with a gridlike code," *Science*, vol. 352, no. 6292, pp. 1464–1468, Jun. 2016.