# arXiv:2503.00612v1 [cond-mat.dis-nn] 1 Mar 2025

# **Statistical Mechanics of Semantic Compression**

Tankut Can\*

Department of Physics, Emory University, Atlanta, GA (Dated: March 4, 2025)

The basic problem of semantic compression is to minimize the length of a message while preserving its meaning. This differs from classical notions of compression in that the distortion is not measured directly at the level of bits, but rather in an abstract semantic space. In order to make this precise, we take inspiration from cognitive neuroscience and machine learning and model semantic space as a continuous Euclidean vector space. In such a space, stimuli like speech, images, or even ideas, are mapped to high-dimensional real vectors, and the location of these embeddings determines their meaning relative to other embeddings. This suggests that a natural metric for semantic similarity is just the Euclidean distance, which is what we use in this work. We map the optimization problem of determining the minimal-length, meaning-preserving message to a spin glass Hamiltonian and solve the resulting statistical mechanics problem using replica theory. We map out the replica symmetric phase diagram, identifying distinct phases of semantic compression: a first-order transition occurs between lossy and lossless compression, whereas a continuous crossover is seen from extractive to abstractive compression. We conclude by showing numerical simulations of compressions obtained by simulated annealing and greedy algorithms, and argue that while the problem of finding a meaningpreserving compression is computationally hard in the worst case, there exist efficient algorithms which achieve near optimal performance in the typical case.

### I. INTRODUCTION

Human working memory has a limited capacity, as revealed from numerous experiments using unstructured stimuli [1]. Nevertheless, we have the ability to process information on extremely long timescales, in apparent contradiction to this finite capacity. The crucial ingredient for accomplishing this is compression, often called "chunking" in cognitive science, whereby stimuli are recoded into more compact representations [2].

Compression is routinely observed in social communication. Bartlett [3] showed that when stories are transmitted between humans, they tend to become shorter and more stereotyped. Furthermore, in experiments on human memory for narratives, in which subjects read a story and are subsequently asked to retell it, there is a strong tendency to produce a compressed version of the story in the retelling, using efficient paraphrases and summaries [4-6].

Importantly, human communication involves lossy compression. For instance, in the narrative memory experiments described above, the verbatim text of the original story, also known as the surface structure, cannot be reconstructed from the recall of participants. If the surface structure does not seem to matter, then what is being transmitted during communication? A natural hypothesis is that *meaning* is the important thing, and surface structure can be sacrificed as long as meaning is kept invariant.

But what is meaning? This notoriously elusive concept finds its most concrete formulation in the study of semantics, which seeks to understand how meaning arises in language [7]. Far from being a mere formal exercise, the study of semantics is central to the psychology of human memory. Short-term sentence recognition experiments show that details of the wording of a sentence are easily forgotten, while the meaning or "gist" of a sentence is kept much longer and more stably in memory [8-10]. This can be seen by testing paraphrases which preserve meaning, against sentence variants which change the original meaning. Measuring forgetting over longer timescales (days to months) confirms this observation, showing that memory for surface structure decays much faster than memory for semantic structure, which includes higher-level abstractions of a text that together give it meaning to an individual [11, 12]. In short, our memory for discourse appears to be primarily semantic in nature. Therefore, while the compression involved in human communication may lose surface structure, it tends to preserve semantic structure; for this reason, we refer to this process as *semantic compression*.

Semantic compression therefore plays a central role in human communication. Traditionally, lossy compression is the purview of rate-distortion theory, and semantic compression has been studied precisely in this context [13, 14]. Furthermore, there have been more general theories which frame pragmatic communication between agents in terms of optimal transport [15]. While necessary, these general theoretical frameworks leave open some questions about the mechanism of compression in particular settings. For instance, what is the interplay between the structure of representations in semantic space, and the capacity for compression? To address such a question, we must try to make explicit contact with the representations that humans (and machines) make use of. In other words, we must specify the semantic distortion function.

To define the distortion function, we take inspiration

<sup>\*</sup>Electronic address: tankut.can@gmail.com

from both machine learning and cognitive neuroscience. Lexical items, such as words and phrases, are clearly stored in long-term memory [7]. Studying brain responses during a story listening task, researchers were able to map out what they called a semantic network in the brain [16, 17]. This was then used to argue that semantic representations are continuously represented in brain activity in the whole cortex [18].

The idea of semantic spaces has a long history in psychology, dating back to the spreading activation theory in the late 60's [19], and later [20]. Similarly, Gärdenfors [21] makes the argument for continuous conceptual spaces in neural population representations. Semantic spaces were also studied in [22], and argued to be closely related to the brain's native instruments for representation of spatial geometry. This mirrors the ability of hippocampus to generate cognitive maps which represent abstract categories instead of spatial location [23]. The question remains: how does the brain represent the presumably high-dimensional spaces involved in our semantic knowledge base? Indeed, it was been argued that low-dimensional spaces are insufficient to describe the geometry of concepts [24], and that complex networks [25] or high-dimensional distributed representations [26] are needed to account for semantic similarity judgements. A recent review has even argued in favor of explicit vector space representations of concepts [27].

In parallel research, language modeling in machine learning has naturally come upon the idea of using continuous vector spaces to represent word meanings. Some particularly vivid examples of this come from algorithms such as word2vec [28] or GloVe [29], which map each word or token in a lexicon to an *embedding vector* which lives in a high-dimensional Euclidean space. Remarkably, meaningful relations between words, such as analogies, are found to be encoded in geometric relations between their vector embeddings. For instance, the vector sum of 'royal' and 'man' is close to 'king' and 'prince' (albeit also close to other seeming non-sequiturs). Similarly, analogies can be represented by vector addition, as with,  $\mathbf{v}(\text{duke}) - \mathbf{v}(\text{male}) \approx \mathbf{v}(\text{duchess}) - \mathbf{v}(\text{female})$ . More generally, contrastive representation learning is an approach in deep learning that seeks to capture semantic similarity between any inputs (words, sentences, images) in an embedding space using continuously differentiable embeddings [30]. In a real sense, deep learning rests on the power and efficacy of a continuous semantic space.

Given the stunning success of language modeling, and the evidence from cognitive neuroscience, we assume that the space of meaning, or semantic space, is given by a Euclidean vector space. Furthermore, we will assume that two meanings are similar if they are close in Euclidean distance in this semantic space. Every message, no matter its length, will be represented in this semantic space by a vector. We further assume a "bag of words" representation for every message, in which the embedding of a long message is just the linear sum of the embeddings of all of its constituent lexical items. This allows us to define the problem of semantic compression as one which minimizes the Euclidean distance between the embeddings of two messages, subject to the constraint that the compressed message has fewer constituent tokens than the target message. All of these objects will be defined mathematically below.

In this paper, we introduce a statistical mechanical model for semantic compression, and present its meanfield phase diagram under the replica symmetric (RS) Ansatz. We show how the RS theory reveals that qualitatively different phases of compression are encountered as one varies embedding dimension D, lexicon size N, target message length L, and compressed length  $\overline{L}$ . In particular, we identify a transition between lossy and lossless compression, which we conjecture is a first-order transition. Furthermore, in the lossy phase, we can identify regimes, related by a crossover behavior, in which the compression is either: 1) extractive, wherein the target can only be compressed by removing words or tokens; 2) abstractive, in which multiple words in the target message are represented by a single word that does not appear in the target. We compare our phase diagram and order parameters to numerical experiments, showing where agreement is good, and in which regimes agreement breaks down. RS theory describes extractive lossy compression very well, but fails at capturing the details of the transition to lossless compression. Finally, we compare a costly Monte Carlo minimization of the distortion, to an efficient greedy algorithm that minimizes the distortion one token at a time by always finding the next closest token embedding. Remarkably, for the typical case scenario we study in this paper, the greedy algorithm is nearly optimal, and finds solutions that are well described by RS MFT in certain regions of the phase diagram.

### II. MATHEMATICAL MODEL OF SEMANTIC COMPRESSION

Here we lay out a set of simplifying assumptions that will allow us to introduce a tractable model of semantic compression.

Assumption 1: The semantic space is Euclidean space  $\mathbb{R}^{P}$ .

Humans have at their disposal a large lexicon of words, word fragments, and phrases, all of which are stored in long-term memory and used in a combinatorial manner (e.g. via grammar) to construct messages. We denote the lexicon with N items by  $\mathcal{L}_N$ , and denote the lexical items or listemes [7] by  $s_i \in \mathcal{L}$ , i = 1, ..., N. A message of length L is defined as a sequence of L lexical items

$$S(\mathbf{k}) = s_{k_1} s_{k_2} \dots s_{k_L}, \quad |S| = L.$$
(1)

From here we see that every message can be represented by a vector  $\mathbf{k} \in \mathbb{Z}_N^L$ , where  $k_1$  is the integer label of the first listeme in the message, and so forth. It is useful to represent a message by a *count vector*  $c(S) \in \mathbb{Z}^{+N}$ , where each entry  $c_i$  is a positive integer that gives a count of the number of times listeme  $s_i$  appears in S:

$$c_i = \sum_{j=1}^L \delta_{ik_j}.$$
 (2)

Henceforth, we drop the argument of c = c(S) and simply refer to vector c as the message.

Every message has a representation in semantic space  $\mathbb{R}^{P}$ . If we denote the space of all messages S, then we define an *semantic embedding* function which maps every message to a point in semantic space:

$$X: \mathcal{S} \to \mathbb{R}^P. \tag{3}$$

We assume that each individual lexical item has a unique semantic embedding  $X(s_i) = E_i$ . To make progress with our model, we make the next crucial simplifying assumption

Assumption 2: The semantic embedding of a message is a linear sum of the embeddings of each constituent lexical item, i.e.

$$X(S(\mathbf{k})) = \sum_{j} X(s_{k_j}) = \sum_{j=1}^{L} E_{k_j} = \sum_{i=1}^{N} c_i E_i.$$
 (4)

After the last equality, we have given a representation in terms of the count vector. We primarily use this representation in the rest of the paper.

This assumption states that the meaning of an item in a message is independent of the structure of the message, i.e. independent of context. In natural language processing, this representation is usually referred to as a "bagof-words", since the embedding only depends on the set of words or listemes used, and is insensitive to the order or general context in which they are used. For example, with a bag-of-words representation, we cannot semantically distinguish "Dog bites man" from "Man bites dog", since these are composed of the same lexical elements. It is of course possible to expand the lexicon to include compound phrases like "(dog, subject)" as well as "(dog, agent)", but this would potentially lead to an unbounded growth of the lexicon, due to combinatorial explosion. Indeed, this endeavor amounts to capturing the infinite generative power of syntax with a fixed lexicon, which seems both unfeasible and extremely inelegant. Therefore, we fully acknowledge that our embedding has some obvious shortcomings, but nevertheless pursue the consequences to the end. We will find that even with this simple choice, there is a rich structure to semantic space. Inclusion of syntax and context more generally must be reserved for future work.

The next assumption we make concerns the metric of semantic similarity.

Assumption 3: We take semantic dissimilarity, or equivalently semantic distortion, between two messages to be quantified by the Euclidean distance squared between their embeddings,

$$d(S, S') = ||X(S) - X(S')||^2.$$
 (5)

A small distortion arises when semantic embeddings are close in semantic space. The most straightforward generalization of this entails imbuing the semantic space with a nontrivial metric. For instance, it has been argued that olfaction utilizes a hyperbolic embedding space [31]. In many pre-trained machine learning embeddings (e.g. OpenAI, SBert, etc.), the semantic space is the unit P-sphere, in which case the distortion function is equivalent to the cosine similarity  $\cos(S, S')$  after a shift, i.e.  $d(S, S') = 2 - 2\cos(S, S')$ .

The preceding assumptions concern the structure of semantic space and semantic similarity. The next few assumptions concern the structure of embeddings and the messages.

Assumption 4: The vector embeddings  $E_i$  are random Gaussian vectors with the following moments

$$\mathbf{E}\left[E_{i}^{\mu}\right] = b^{\mu}, \quad \mathbf{E}\left[(E_{i}^{\mu} - b^{\mu})(E_{j}^{\nu} - b^{\nu})\right] = \delta_{\mu\nu}\Sigma_{ij}. \tag{6}$$

The finite mean value affects all embeddings in essentially the same way. However, this uniform shift will have consequences on the overall structure of the embeddings, and consequently on the likelihood of finding efficient compressions or paraphrases. The nontrivial variance is supposed to reflect the fact that embeddings of semantically similar lexical items will tend to be correlated. Thus, under an appropriate indexing of the lexicon,  $\Sigma_{ij}$ is expected to have a block diagonal structure. In the main discussion below, we specialize to the case of uncorrelated embeddings  $\Sigma_{ij} = \delta_{ij}$ , with zero mean  $b^{\mu} = 0$ . We explore the general setting in an upcoming paper.

The assumption of random embeddings is not unreasonable. In fact, it can be observed that word embeddings from popular algorithms (e.g. word2vec or GloVe) appear to have components which follow a Gaussian distribution. These algorithms assume randomly initialized vectors assigned to each word, which are subsequently updated by some learning rule. It it conceivable that this learning rule does not change the distribution, but the relative position of the random vectors. This means that any collection of random vectors can be used as embedding vectors - the only question is which word gets assigned to which vector. This fascinating argument was made in [32], but for random points in hyperbolic space instead of high-dimensional Euclidean space.

Finally, we constrain the space of messages:

Assumption 5: The components if the count vector defined in Eq. 2 are binary, i.e.  $c_i \in \{0, 1\}$ .

This means messages are not permitted to have any repeated lexical items  $s_i$ . If we treat the lexical items as words in an actual text, then this assumption is obviously wrong. However, we may treat the lexicon not as representing individual words but unique concepts. In this case, it is a little more sensible to have a message that does not have repeating concepts.

Having laid out the essential details of the semantic space and the embedding function, we are now in a position to define a statistical mechanics of semantic compression.

# III. STATISTICAL MECHANICS OF SEMANTIC COMPRESSION

We will ultimately be interested in the thermodynamic scaling limit, which involves taking the lexicon size to infinity. Since for our bag-of-words embedding function, each message is represented uniquely by the vector of counts c, we henceforth denote the original message by c, and the compressed message by  $\bar{c}$ . The 1-norm of the counts vector gives the total length of the message, which we denote as  $||c||_1 = L$  and  $||\bar{c}||_1 = \bar{L}$  for the original and compressed message, respectively. The Hamiltonian is defined to be

$$H(c,\bar{c}) = \frac{1}{2N}d(c,\bar{c}) = \frac{1}{2N}\sum_{i,j}\sigma_i J_{ij}\sigma_j, \qquad (7)$$

where  $\sigma_i = c_i - \bar{c}_i$ , and  $J_{ij} = E_i \cdot E_j \equiv \sum_{\mu=1}^{P} E_i^{\mu} E_j^{\mu}$ . With this Hamiltonian we can proceed to define a statistical mechanics formulation of our combinatorial optimization problem [33]. For quenched embeddings  $E_i$  and original message c, we must find the optimal compression  $\bar{c}$ of a fixed length  $\bar{L}$ . Therefore, we define the partition function

$$Z(E,c,\bar{L}) = \sum_{\bar{c}_i, \|\bar{c}\|_1 = \bar{L}} \exp\left(-\beta H(c,\bar{c})\right), \qquad (8)$$

where the sum is constrained to be over all messages  $\bar{c}$  of a fixed length  $\bar{L}$ . From this, we obtain the free energy density

$$f_{\beta}(L,\bar{L},b,\Sigma) = -\frac{1}{\beta N} \mathbb{E}\left[\log Z(E,c,\bar{L})\right]_{E,c,\|c\|_{1}=L}.$$
 (9)

Here again we take a constrained average over original messages c at a fixed length  $||c||_1 = L$ . We also average over random embeddings.

The interesting question that our model is supposed to address is how the structure of embeddings is implicated in the ability to produce efficient semantic compressions. For trivial correlations in Eq. 6,  $\Sigma_{ij} \propto \delta_{ij}$ , the only relevant structure is the dimension of the embedding space P, and the size of the lexicon N. Therefore, we consider how the compression scales with their ratio

$$\alpha = \frac{P}{N},\tag{10}$$

which we refer to as the *relative embedding dimension*. Of course, the size of the original message and the target compressed length will interact with  $\alpha$  to influence compressibility. Therefore, we define here the message length ratios.

$$\ell = \frac{L}{N}, \quad \bar{\ell} = \frac{\bar{L}}{N}. \tag{11}$$

We also introduce the compression ratio

$$C = \bar{L}/L, \tag{12}$$

as an important control parameter in our model. The thermodynamic limit in our model amounts to taking  $P, N, L, \overline{L} \to \infty$  while keeping fixed  $\alpha, \ell$ , and  $\overline{\ell}$ .

The average distortion in our model is just the mean energy density, and is given by

$$D(\beta) = \partial_{\beta} \left(\beta f_{\beta}\right). \tag{13}$$

We can also find the minimum distortion from the zero temperature limit

$$D_{min} = \frac{1}{N} \mathbb{E} \left[ \min_{\bar{c}} H(c, \bar{c}) \right]_{E,c} = \lim_{T \to 0} f_{\beta}.$$
(14)

### IV. ORDER PARAMETERS FROM MEAN-FIELD THEORY

The calculation of the free energy is carried out by straightforward application of replica theory. We present the details in the supplemental material. Below, we give the main results and their interpretation. But first, we introduce the order parameters and provide an intuition for their meaning by studying a simple example. These order parameters are in fact quite natural in the context of spin glasses, but their meaning in the present context is not immediately apparent.

The first order parameter is the overlap

$$R = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}c_i\langle\bar{c}_i\rangle\right]_{E,c},\qquad(15)$$

where we denote by brackets  $\langle ... \rangle$  the average using the partition function (8) (i.e. the thermal average), with  $E_i$  and c fixed (quenched). For binary counts  $c_i, \bar{c}_i \in \{0, 1\}$ , the overlap obeys the bounds

$$\max\left(0,\ell+\bar{\ell}-1\right) \le R \le \bar{\ell}.\tag{16}$$

The upper bound is saturated when the compression is purely **extractive**. This means that the only effective compression possible is one in which a subset of the original lexical items are used. An example of an extractive compression would be if "The quick brown fox jumps over the lazy dog" was shortened to "The fox jumped over the dog". Below this upper bound, the compression must utilize paraphrases, since it would require some of the words in  $\bar{c}$  to not have appeared in c. A paraphrasing compression might look like "A fast fox leaped over the canine". Approaching the lower bound requires an abstractive compression, in which a majority of lexical items in the compression  $\bar{c}$  are not in the original message c. For this well-known example sentence that uses every letter in the English alphabet, an efficient abstractive compression could be "a famous pangram".

There is also an Edwards-Anderson (EA) order parameter [34] characterizing the overlap between different "ground-state" configurations:

$$Q = \bar{\ell} - \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^{N} \langle \bar{c}_i \rangle \langle \bar{c}_i \rangle\right]_{E,c}.$$
 (17)

Within the replica symmetric theory, Q is single-valued and signals a phase transition. The range of the EA order parameter for binary messages is given by

$$0 \le Q \le \min\left(\bar{\ell}, 1 - \bar{\ell}\right). \tag{18}$$

The lower bound Q = 0 is saturated in the case that there is a unique compression  $\bar{c}$ . When there are multiple states which achieve the minimal distortion, then Q > 0. We will see in what follows that for random Gaussian embeddings, Q = 0 corresponds to the lossy compression phase, whereas Q > 0 characterizes the lossless compression phase.

We can gain some intuition for these order parameters by considering first a simple limit of our model for semantic compression.

### A. Special Case: Weighted Hamming Compression

For orthogonal patterns  $E_i \cdot E_j = w_i \delta_{ij}$  (which requires  $P \ge N$ ), the Hamiltonian is the weighted Hamming distance between these bit strings

$$H(c,\bar{c}) = \frac{1}{2N} \sum_{i=1}^{N} w_i (c_i - \bar{c}_i)^2.$$
(19)

The minimal distortion that is achievable is

$$D_{min} = \frac{1}{N} \mathbb{E} \left[ H_{min} \right] = \frac{\langle w \rangle}{2N} (\ell - \bar{\ell}), \qquad (20)$$

which obtains when  $\bar{c}_i$  is only nonzero for *i* such that  $c_i = 1$ . In other words,  $R = \bar{\ell}$  is saturated at its upper bound, and the compressions are strictly **extractive**.

Since each bit is unequally weighted, there will generically be a unique minimizer, which implies Q = 0. This situation is slightly different for the pure Hamming distance which has  $w_i = 1$  for all *i*. In that case, there will be a large degeneracy of minimal distortion compressions, which leads to Q > 0.

We will see below that for random embeddings, the compression phase diagram is described by Hamming compression in the limit that  $\alpha >> 1$ .

### V. REPLICA SYMMETRIC MEAN FIELD THEORY

The replica symmetric mean-field theory (RS MFT) is most naturally formulated in terms of the order parameter

$$q_{ab} = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\sigma_{i}^{a}\sigma_{i}^{b}\right]_{E,c,\bar{c}}.$$
(21)

This can be related in a straightforward way to the more "physical" overlap (15) and EA parameter (17). In particular, for the replica symmetric ansatz,  $q^{ab} = q_0 \delta_{ab} + (q_0 - q)$ , we have

$$q_0 = \ell + \bar{\ell} - 2R, \quad q = q_0 - Q.$$
 (22)

We review the full RS MFT in the supplementary material. For now, we present only the zero temperature ( $\beta \rightarrow \infty$ ) limit. For small temperatures,  $Q = Q_0 + TQ_1 + O(T^2)$ . In the lossy compression phase,  $Q_0 = 0$ ,  $q_0 \rightarrow q \equiv \bar{q}$ , and the RS MFT reduces to a set of two equations for  $\bar{q}$  and an auxiliary (Lagrange multiplier) variable  $\lambda$ :

$$\bar{q} = \ell H_1(-\lambda) + (1-\ell)H_1(\lambda), \qquad (23)$$

$$\ell - \bar{\ell} = \ell H_1(-\lambda) - (1-\ell)H_1(\lambda), \qquad (24)$$

where

$$H_1(\lambda) = \frac{1}{2} \operatorname{erfc}\left(\frac{\alpha/2 + \lambda}{\sqrt{2\alpha\bar{q}}}\right).$$
(25)

The second equation Eq. 24 arises due to the hard constraint on the compression length  $\bar{\ell}$ . The mean distortion is equal to the minimal distortion in this limit, and given by

$$D = \frac{\alpha \bar{q}}{2(1+Q_1)^2},$$
 (26)

where

$$Q_1 = \frac{\ell H_2(-\lambda) + (1-\ell)H_2(\lambda)}{1-\ell H_2(-\lambda) - (1-\ell)H_2(\lambda)},$$
(27)

and  $H_2(-\lambda) = \partial_{\lambda}H_1(-\lambda)$ . For  $Q_0 > 0$ , the zero temperature energy is zero, which implies a zero distortion compression. The mean-field equations are

$$q_0 = \ell F_1(-\lambda) + (1 - \ell) F_1(\lambda), \tag{28}$$

$$q = \ell F_2(-\lambda) + (1-\ell)F_2(\lambda), \qquad (29)$$

$$\ell - \bar{\ell} = \ell F_1(-\lambda) - (1-\ell)F_1(\lambda), \tag{30}$$

where

$$F_k(\lambda) = \int Dz \left(1 + \exp\left(\Theta(z,\lambda)\right)\right)^{-k}, \qquad (31)$$

$$\Theta(z,\lambda) = -\frac{1}{q_0 - q} \left[ \sqrt{\alpha q} z - \frac{\alpha}{2} - \lambda \right], \tag{32}$$

and  $Dz = dz \exp(-z^2/2)/\sqrt{2\pi}$ . Here, Eq.30 comes from the hard constraint on compression length.

We now explore various limits of the mean-field theory.

## A. Phase Diagram

Fig. 1A shows the zero temperature phase diagram obtained from the RS MFT describe above. In general, the order parameters will depend on the parameters  $\alpha, \ell, \bar{\ell}$ ,



FIG. 1: Semantic Compression Phase Diagram and Order Parameters A) The zero temperature phase diagram for the RS order parameters fixing  $\ell = 0.4$ . The discontinuous transition is indicated by the thick black line. In the lossless, compressible phase (blue region), the EA order parameter Q > 0, and the RS MFT has a unique solution with zero mean distortion. Within the lossy phase, where the EA order parameter Q = 0, there is a region (shaded orange, enclosed by black dashed curves) in which the RS MFT has multiple solutions. Outside of this region (white area), the RS MFT has a unique solution. Colored dashed lines show the slice along which the order parameters are computed in the other panels. B) Average distortion (at zero temperature) normalized by its value  $\alpha \ell/2$  at C = 0. Outside the lossless phase, the distortion never reaches zero, but generally decreases with compression ratio. For the green curve, the value of  $\alpha \approx .251$  is chosen such that the distortion hits zero at exactly C = 0.6. C) The overlap order parameters in the lossy phase. For reference, we show the Hamming compression limit (red dashed) in which Q = 0 and  $R = \ell$ . For  $\alpha = 0.5$  (solid blue), there is a bifurcation in the RS MFT at  $C \approx 0.94$ , above which two new solutions appear which are much closer to the Hamming limit. For larger  $\alpha$ , the overlap is close to the Hamming line for all compression ratios. D) Overlap and EA order parameter for  $\alpha \approx 0.251$ . There are no Q = 0 solutions to the order parameters in the region between the two vertical lines at C = 0.6 (solid black) and  $C \approx 0.98$  (dotted black). However, Q > 0solutions actually appear at compression ratios slightly smaller than C = 0.6. There is no reason to prefer one of these over the other in the RS theory, since they both have negative entropy. Furthermore, the Q > 0 solutions survive until C = 1. However, approaching C = 1 the RS MFT does yield a sensible physical solution, which we believe takes over for larger compression ratios.

but only through the following ratios  $\alpha/\ell$ ,  $C = \ell/\ell$ , and  $\ell/(1 - \ell)$ . We work exclusively within the replica symmetric ansatz. The phase boundaries are drawn in the following manner: if there exists a solution to the MFT Eqs. 23 and 24 that has  $Q_1 > 0$ , we assume the system is in the lossy phase. The lossy phase is impossible if

 $Q_1 < 0$ , and we denote that by the condition (at zero temperature) that  $Q_0 = Q > 0$  (shaded blue in Fig. 1A). We have not logically ruled out the possibility that lossless compression happens outside this region.

For simplicity, we fix  $\ell = 0.4$  to be able to visualize the phase diagram as  $\alpha$  and C are varied. For a given  $\ell$ , there is a maximal  $\alpha^*$  above which lossless compression is impossible. This corresponds to the point where the solid black line meets the dashed black line, indicated by a star in Fig.1A. We find numerically that  $\alpha^*(\ell)$  is approximately quadratic in the interval [0,1], and goes to zero at the boundaries. The peak obtains at  $\ell \approx 0.53$ , with  $\alpha^*(\ell^*) \approx 0.4049$  (see Fig. 3 in the supplemental material).

### B. Small Lexicon Limit

The limit  $\alpha \rightarrow 0$  is somewhat trivial in our model, due to the explicit scaling we use for the Hamiltonian. Nevertheless, it is instructive to describe this limit. The RMFT yields

$$q_0 = \ell + \bar{\ell} - 2\ell\bar{\ell}, \quad q = \ell - 2\ell\bar{\ell} + \bar{\ell}^2, \tag{33}$$

From which we get the order parameters

$$R = \ell \bar{\ell}, \quad Q = \bar{\ell} (1 - \bar{\ell}). \tag{34}$$

Furthermore, in this limit, the minimum distortion is precisely zero. This is simply a consequence of the fact that every possible  $\bar{c}$  will produce a distortion that scales like P/N, and thus tends to zero in the thermodynamic limit if P = O(1).

### C. Large Embedding Dimension

We can consider also the limit  $\alpha \to \infty$ , in which the embedding dimension becomes much larger than the size of the lexicon. In this limit, the embeddings become approximately orthogonal, and we expect to recover the weighted Hamming phase. This can be observed directly from the zero temperature RS MFT with  $Q_0 = 0$ , in which we get

$$\bar{q} \approx \ell - \bar{\ell}, \quad D \sim \frac{\alpha(\ell - \bar{\ell})}{2}.$$
 (35)

This follows from 20 by noting that  $\langle w \rangle = P$  for random Gaussian embeddings.

### **D.** Signal Recovery Limit: C = 1

In the limit that the compression ratio C = 1, our model is formally very close to the compressed sensing problem of signal reconstruction studied in [35, 36], except for the fact that our signal and message are binary variables. In this limit, the RS solution has positive entropy (unlike seemingly all solutions with C < 1). For  $\alpha/\ell \geq 1.3257$ , the only solution to Eqs. 23 and 24 is  $\bar{q} = 0$ , which makes D = 0 and thus corresponds to perfect signal recovery. For smaller  $\alpha/\ell$ , the recovered signal will not be perfect. This regime falls inside the orange shaded region enclosed by the dashed black lines in Fig. 1A. where Q = 0.

### E. Comparison to numerics

It is important to note that in the zero temperature limit, the entropy is strictly negative throughout the phase diagram, except for compression ratios very close to one (inside the orange shaded region in Fig.1A. On the surface, this would make our phase diagram meaningless. However, comparing to numerics, we find that the phase diagram is at the very least qualitatively accurate, if not quantitatively correct.

In Fig. 2, we compare the theoretical results to two optimization algorithms: simulated annealing (SA) and a greedy algorithm (GA). It is quite interesting to note that in most regions of the phase diagram, the greedy algorithm is significantly faster and performs nearly as well as, if not better than, SA in minimizing distortion. Perhaps the most interesting observation is that in the lossless compression phase, the GA is well approximated by the RS MFT. Overall, we observe that the RS MFT offers decent predictions for large  $\alpha$  (approaching the Hamming phase), and for small compression ratio C. However, approaching the region with Q > 0, the theory apparently breaks down. We attribute this to both the patently wrong approximation of replica symmetry and finite-size effects in the numerical simulations (which we performed for rather small systems).

# VI. DISCUSSION

We have introduced and solved a statistical mechanics model of semantic compression. In this work, we have learned that even with completely random embeddings, semantic compression undergoes a phase transition between lossy and lossless compression. The detailed structure of the phase diagram is governed both by properties of the semantic space (relative embedding dimension  $\alpha$ ), as well as the compression ratio. We have also found crossover behavior between extractive and abstractive summarization.

While this work has focused on formulating the mathematical problem and solving it in the mean field limit, it raises tantalizing questions about fitting to real-world language data. For instance, where does typical communication lie on this phase diagram? We may speculate using some details from modern language models. Typical lexicon size is of order  $10^4$ , while embedding dimensions are  $P \sim O(10^2)$ , giving  $\alpha \sim 0.01$ . At this value, the compression ratio does not have to be very large before reaching the phase boundary. Across this phase boundary, there will tend to be many summaries which are very good. In other words, for any given message c, it will be possible to find many paraphrases which roughly mean the same thing. This is, at least intuitively, precisely the situation with natural language. Mathematically, this means the probability distribution of a message conditioned on its meaning, P(c|M), has nonzero entropy. A similar quantity, referred to as the "wording



FIG. 2: Numerics Comparing numerical optimization via simulated annealing (SA) and a greedy algorithm (GA), with RS MFT. A) The SA values (crosses) of the order parameters diverge from RS MFT prediction (solid curves) in the regime we expect to see the phase transition to lossless and abstractive compression (for the green points, past C = 0.6). There is also a striking deviation for the blue ( $\alpha = 0.5$ ) for intermediate C, but agreement for small C and  $C \rightarrow 1$ . Curiously, when disagreement between RS MFT and SA is large, the GA finds solutions with an order parameter that is very close to that predicted by the RS MFT. B) The numerically computed EA order parameter shows a smooth rise and fall, in contrast to the theoretical prediction. Since the GA produces a unique minimizer for a given c, there is no comparison to be made here with SA, which generally finds multiple minimizers for a given quenched message. C) The average distortion is fairly well described by theory, with notable deviations for larger compression ratios, as in the previous plots. We used N = 200 and average over 10 embedding (disorder) realizations. The order parameters are computed by estimating the low energy spectrum of the Hamiltonian and computing a truncated temperature average (see Appendix A for details)

information", was recently shown to be nonzero using large language models [37]. Estimating  $\alpha$  is more difficult for humans. We may estimate a lexicon size which is roughly comparable in order of magnitude [38], but it is anyone's guess what the dimension of a human's semantic space may be. For instance, the fMRI study of [18] found that the semantic space should have at least four dimensions, setting a very small lower bound on  $\alpha$ . Other sources seem comfortable with semantic (or conceptual) spaces with a huge number of dimensions, resulting in a large relative embedding dimension [27]. Of course, the structure of real semantic embeddings is not random, and the joint embedding of a message is likely not additive but context dependent. It would be interesting to explore how these properties influence the compressibility of a language. For instance, structured embeddings might allow for meaning-preserving compression even at very large  $\alpha$ . We leave these tantalizing questions for future work.

Optimization and Mixed Integer Linear Programming The process of semantic compression we have studied in this paper is an example of a mixed-integer linear programming problem. This tells us that the problem is in fact NP-hard. The remarkable fact is that we can find a fast algorithm which is  $O(\bar{T}L)$  that can give a good solution to this problem. This is not terribly surprising, since our formulation of semantic compression is very similar to the partition problem, which has been called the "easiest" NP-hard problem [39, 40]. We showed that a simple greedy algorithm often gives a very low distortion. Furthermore, we showed numerically that the greedy algorithm gives a solution that is well described by the replica symmetric order parameters.

Communication and an Alignment Problem: Communication is rife with misunderstanding, and our theory actually sheds some light on a potential mechanism. Suppose Alice sends a message  $c_A$  to Bob, who then produces a compressed message  $\bar{c}_B$  which, according to Bob, is distortion minimizing. Bob chooses his compression using his personal semantic embeddings. Now Alice can compare  $\bar{c}_B$  to  $c_A$ , by either computing the overlap or the distortion. But to calculate distortion, Alice can only use her own embeddings. If Alice and Bob have perfectly aligned semantic embeddings, then the message  $\bar{c}_B$  would presumably look like a compression that Alice could have come up with. Therefore, she will agree that  $\bar{c}_B$  means the same thing as  $c_A$ , and will conclude that Bob understood the original message. However, if the embeddings are not aligned, the distortion will increase with the degree of misalignment. Therefore, the simple fact that Alice recognizes semantic similarity between her original message, and Bob's repeated version of her message, implies some degree of alignment between their separate, private, semantic spaces. Note also that the crucial thing is not that they both have vectors pointing in the same direction - if all of Bob's embeddings are related to Alice's embeddings by the same orthogonal transformation, then although their embedding vectors might be different, the Hamiltonian, and hence distortion function, is unchanged. So really, the crucial property is the relative positions of the embeddings. Surprisingly, there is experimental support for shared semantic dimensions [18]. But after thinking through the problem of communication, we conclude that such a shared space is hardly surprising at all, and in fact is necessary for people speaking a common language.

In fact, the thought experiment above illustrates a general semantic alignment problem: does communication between individuals require an alignment between their semantic spaces? In the most general setting, we might replace alignment with isomorphism, especially if the semantic spaces are mathematically different spaces. For the model of semantic compression we consider in this paper, the notion of alignment is taken from linear algebra. However, this formulation extends beyond the realm of human communication. Suppose Bob is a large language model (LLM), which are known to be excellent summarizers [43]. The scenario above resembles modern variants of Turing's imitation game [41], wherein Bob simply needs to convince Alice that he is human. In this semantic compression game, Bob simply needs to convince Alice that he *understands* her, by doing what any good student does: summarizing what Alice says in his own words, but without losing the meaning of the original message. And if Bob is an LLM, then in all likelihood he will win this game. We conclude that for this to be possible, there must exist an isomorphism between the semantic space constructed by our brains, and the latent representations utilized by LLMs in performing their computation. We speculate that this mathematical isomorphism between semantic spaces is at the root of what is referred to as "common ground" in linguistics and philosophy, which encompasses the knowledge base shared between individuals that provides the scaffolding for effective communication [42].

In summary, we have formulated semantic compression as a combinatorial optimization problem for a spin glass Hamiltonian, and solved the statistical mechanics in the replica symmetric limit. We find that for a lexicon randomly embedded in semantic space, the compressibility of a message undergoes both phase transitions and crossover behavior, as a function of embedding dimension, message length, and compression ratio. For sufficiently small embedding dimension, small compression ratios tend to incur distortion, whereas larger compression ratios are lossless. Furthermore, the compressed messages in the lossless phase tend to be abstractive, using lexical items outside the original message for more efficient summaries. For larger embedding dimension, there is no phase transition, and compression always incurs a cost. In this region of the phase diagram, the compressed messages tend to be extractive, restricted to using lexical items from the original message. Finally, we show that while the original optimization problem falls in the class of mixed-integer linear programming, and is therefore NP-hard, we were able to find an efficient greedy algorithm that is competitive with the more costly simulated annealing.

We assumed semantic embeddings had no correlations, which probably does not reflect the structure of such embeddings in the wild. We will examine the influence of structure and correlation in semantic space in future work. Our theory also has introduced two novel order parameters, the overlap and an EA order parameter, which can be applied to the study of language. We hope to explore this in follow-up work.

Acknowledgements I have benefited from numerous conversations on this topic, and would like to thank Mikhail Katkov, Misha Tsodyks, Weishun Zhong for their feedback. I am especially indebted to Bruno Loreiro and Francesca Mignacco for some critical insights at the initial stages. I acknowledge the support of the Eric and Wendy Schmidt Membership in Biology, the Simons Foundation, and the Starr Foundation Member Fund in Biology at the Institute for Advanced Study, where most of this work was completed.

- [1] N. Cowan, *Working memory capacity* (Psychology press, 2012).
- [2] G. A. Miller, Psychological review **63**, 81 (1956).
- [3] F. C. Bartlett, *Remembering* (Cambridge University Press, 1932).
- [4] E. Musz and J. Chen, Communications Biology 5, 1 (2022), ISSN 23993642.
- [5] A. Georgiou, T. Can, M. Katkov, and M. Tsodyks, Learning & Memory **32**, a054043 (2025).
- [6] W. Zhong, T. Can, A. Georgiou, I. Shnayderman, M. Katkov, and M. Tsodyks, bioRxiv pp. 2024–12 (2024).
- [7] R. Jackendoff, Foundations of Language: Brain, Meaning, Grammar, Evolution (Oxford University Press, 2002).
- [8] B. R. Gomulicki, Acta Psychologica 12, 77 (1956), ISSN 00016918.
- [9] J. S. Sachs, Perception and Psychophysics 2, 437 (1967).
- [10] S. Fillenbaum, Language and Speech 9, 217 (1966).
- [11] W. Kintsch, D. Welsch, F. Schmalhofer, and S. Zimny,

Journal of Memory and language **29**, 133 (1990).

- [12] W. Kintsch, Comprehension: A paradigm for cognition (Cambridge university press, 1998).
- [13] B. Guler and A. Yener, in 2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS) (IEEE, 2014), pp. 431–436.
- [14] J. Liu, W. Zhang, and H. V. Poor, IEEE International Symposium on Information Theory - Proceedings 2021-July, 2894 (2021), ISSN 21578095, 2105.04278.
- [15] P. Wang, J. Wang, P. Paranamana, and P. Shafto, Advances in Neural Information Processing Systems 33, 17582 (2020).
- [16] J. Tang, A. LeBel, S. Jain, and A. G. Huth, Nature Neuroscience 26, 858 (2023).
- [17] A. A. Kumar, Psychonomic Bulletin & Review 28, 40 (2021).
- [18] A. G. Huth, S. Nishimoto, A. T. Vu, and J. L. Gallant, Neuron 76, 1210 (2012), ISSN 08966273.

- [19] M. R. Quillian, Behavioral science 12, 410 (1967).
- [20] A. M. Collins and E. F. Loftus, Psychological review 82, 407 (1975).
- [21] P. Gardenfors, The Geometry of Meaning: Semantics based on conceptual spaces (MIT press, 2014).
- [22] S. Viganò, V. Rubino, A. Di Soccio, M. Buiatti, and M. Piazza, NeuroImage 232, 117876 (2021).
- [23] D. Aronov, R. Nevers, and D. W. Tank, Nature 543, 719 (2017).
- [24] A. Tversky and J. Hutchinson, Psychological review 93, 3 (1986).
- [25] M. Steyvers and J. B. Tenenbaum, Cognitive science 29, 41 (2005).
- [26] S. Bhatia, R. Richie, and W. Zou, Current Opinion in Behavioral Sciences 29, 31 (2019).
- [27] S. T. Piantadosi, D. C. Y. Muller, J. S. Rule, K. Kaushik, M. Gorenstein, E. R. Leib, and E. Sanford, Trends in Cognitive Sciences 28, 844 (2024).
- [28] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, NIPS pp. 1389–1399 (2017), 1606.08359.
- [29] J. Pennington, R. Socher, and C. D. Manning, EMNLP (2014).
- [30] P. H. Le-Khac, G. Healy, and A. F. Smeaton, IEEE Access 8, 193907 (2020).
- [31] Y. Zhou, B. H. Smith, and T. O. Sharpee, Science advances 4, eaaq1458 (2018).
- [32] S. Nurmukhamedov, T. Mach, A. Sheverdin, and Z. Assylbekov, arXiv preprint arXiv:2204.12481 (2022).

- [33] M. Mézard, G. Parisi, and M. A. Virasoro, Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications, vol. 9 (World Scientific Publishing Company, 1987).
- [34] S. F. Edwards and P. W. Anderson, Journal of Physics F: Metal Physics 5, 965 (1975).
- [35] S. Ganguli and H. Sompolinsky, Physical review letters 104, 188701 (2010).
- [36] M. Vehkaperä, Y. Kabashima, and S. Chatterjee, in 21st European Signal Processing Conference (EUSIPCO 2013) (IEEE, 2013), pp. 1–5.
- [37] D. Sivan and M. Tsodyks, arXiv preprint arXiv:2411.12728 (2024).
- [38] M. Brysbaert, M. Stevens, P. Mandera, and E. Keuleers, Frontiers in psychology 7, 1116 (2016).
- [39] R. E. Korf, IJCAI International Joint Conference on Artificial Intelligence pp. 538–543 (2009), ISSN 10450823.
- [40] B. Hayes, American Scientist **90**, 113 (2002).
- [41] A. M. Turing, Mind **59**, 433 (1950).
- [42] X. Hao, Y. Jhaveri, and P. Shafto, Advances in Neural Information Processing Systems 36, 5211 (2023).
- [43] The proof of the pudding is in the eating: when prompted with the request to "Summarize the following text in one sentence: " and then given the previous paragraph, OpenAI's GPT-4 produced the following: "Miscommunication arises from misaligned semantic embeddings between individuals, necessitating a shared semantic space for effective communication."

# **Appendix A: Numerical Simulations**

For the numerics in Fig.2, we searched for compressions with simulated annealing [33]. For a given quenched message c, we initialized the search by taking a random subset of  $\overline{L}$  entries which were equal to one, and setting the rest of the entries to zero. This provides  $\overline{c}_0$ . A stochastic update is made on a state  $\overline{c}_t$  by randomly flipping two bits from  $0 \rightarrow 1$  and  $1 \rightarrow 0$ , to give a new state  $\overline{c}_t^*$  with the same norm. The update rule for the compression  $\overline{c}_t$  is probabilistic, and given by

$$\bar{c}_{t+1} = \begin{cases} \bar{c}_t^*, & \text{w/ probability } P_t \\ \bar{c}_t, & \text{w/ probability } 1 - P_t \end{cases}, \quad P_t = \exp\left(-\gamma_t (H(c, \bar{c}_t^*) - H(c, \bar{c}_t))\right). \tag{A1}$$

Here, H(c, c') is the Hamiltonian defined in the Eq. 7, and  $\gamma_t$  is the inverse temperature of the Monte Carlo algorithm, which we update according to the annealing schedule:

$$\gamma_{t+1} = \begin{cases} r\gamma_t, & \text{if } H(c, \bar{c}_{t+1}) < H(c, \bar{c}_t) \\ \gamma_t, & \text{otherwise} \end{cases}$$
(A2)

For our simulations, we used  $\gamma_0 = 0.04$ , and r = 1.01. After a fixed number of steps (we chose 3000), we stop the algorithm. Running this many times gives an ensemble of  $\bar{c}_a$  from we define an empirical probability with a given  $\beta$  that we will eventually take to be large:

$$p_e(\bar{c}_a) = \frac{e^{-\beta H(\bar{c}_a)}}{\sum_b e^{-\beta H(\bar{c}_b)}}.$$
(A3)

We suppress the c argument of the Hamiltonian, since this variable is fixed throughout. With this empirical measure, we compute the order parameters

$$\hat{q}_{0}(\beta) = \sum_{a} p_{e}(\bar{c}_{a}) \frac{1}{N} ||c - \bar{c}_{a}||_{1}, \quad \hat{q}(\beta) = \sum_{a,b} p_{e}(\bar{c}_{a}) p_{e}(\bar{c}_{b}) \frac{1}{N} (c - \bar{c}_{a}) \cdot (c - \bar{c}_{b}).$$
(A4)

From these we get the empirical overlap and EA order parameter using

$$R_e = \frac{1}{2} (\ell + \bar{\ell} - \hat{q}_0), \quad Q_e = \hat{q}_0 - \hat{q}.$$
(A5)

For Fig.2, we plotted the numerical order parameters using  $\beta = 10$ .

# 1. Greedy Algorithm

The greedy algorithm discussed in the main text proceeds as follows: define the initial vector as the message vector embedding

$$X_0 \equiv X(c) = \sum_{i=1}^{N} c_i E_i.$$
 (A6)

Next, find the lexical item closest to  $X_0$ ,

$$i_1 = \underset{k}{\operatorname{argmax}} \|X_0 - E_k\|^2.$$
 (A7)

Then, update the vector by subtracting this closest lexical item:

$$X_1 = X_0 - E_{i_1}.$$
 (A8)

For a general step, the updating of the vector is

$$i_t = \underset{k}{\operatorname{argmin}} ||X_{t-1} - E_k||^2, \quad X_t = X_{t-1} - E_{i_t}$$
(A9)

Finally, to get the desired length of compression, terminate the process after finding  $i_{\bar{L}}$ . The compressed message then has  $\bar{c}_{i_t} = 1$  for  $t = 1, ..., \bar{L}$ .

For Fig. 2, we employ the greedy algorithm for systems with L = 1000, and obtain all the curves with quenched embeddings, averaging over 500 different random target messages.

### Appendix B: Replica Theory for Semantic Compression

**N.B.** The notation in this appendix is slightly different from the main text. However, there is a simple dictionary, which we give here to hopefully avoid any confusion:

Let the fun begin:

The energy function is given by the Euclidean square distance between the original phrase embedding X(s) and the paraphrase  $X(\bar{s})$ .

$$H = \frac{1}{2N} \|X(s) - X(\bar{s})\|^2 = \frac{1}{2N} \sum_{i,j=1}^N (c_i - \bar{c}_i) E_i \cdot E_j (c_j - \bar{c}_j) \equiv \frac{1}{2N} \sum_{\mu=1}^P (\sigma \cdot E^{\mu})^2, \quad \sigma = c - \bar{c}.$$
 (B1)

The partition function that we compute will include the constraint that  $\|\bar{c}\|_1 = \bar{L}$ , i.e. the one-norm of the paraphrase configuration is fixed. Furthermore, in this problem we are treating the original phrase c and the embeddings E as quenched variables. The partition function at temperature  $T = \beta^{-1}$  is then

$$Z_{\beta}(c, E, \bar{L}) = \sum_{\bar{c}_{i}} e^{-\beta H} \delta(\|\bar{c}\|_{1} - \bar{L}) = \sum_{\bar{c}} \int d\bar{x} \exp\left[-\frac{\beta}{2N} \sum_{\mu=1}^{P} (\sigma \cdot E^{\mu})^{2} + i\bar{x} \left(\bar{L} - \sum_{i} \bar{c}_{i}\right)\right].$$
(B2)

The free energy density which follows from partition function is

$$f_{\beta}(c, E, \bar{L}) = -\frac{1}{\beta N} \log Z_{\beta}(c, E, \bar{L}).$$
(B3)

It is not immediately obvious that this quantity should be self-averaging, and we could in principle consider the distribution of the free energy density. But for simplicity, we take the mean value.

$$f_{\beta}(L,\bar{L},\theta) = -\frac{1}{\beta N} \frac{\sum_{c} \delta(||c||_{1} - L)}{\mathcal{N}(L)} \int dP(E) \log Z_{\beta}(c,E,\bar{L}) \equiv -\frac{1}{\beta N} \langle \log Z_{\beta}(c,E,\bar{L}) \rangle_{c,E}, \tag{B4}$$

where  $\mathcal{N}(L) = \sum_{c} \delta(||c||_1 - L)$  is a normalization factor for the distribution of c which counts the number of the distinct phrases of length L. We also denote by  $\theta$  the set of all parameters that characterize the distribution over embeddings E. We assume this expression for the free energy follows from the zero replica limit

$$f_{\beta}(L,\bar{L},\theta) = -\frac{1}{\beta N} \lim_{n \to 0} \frac{1}{n} \left( \langle Z_{\beta}^{n}(c,E,\bar{L}) \rangle_{c,E} - 1 \right).$$
(B5)

From the free energy, we are interested in the ground state energy, which in our problem has the interpretation of the minimal distortion paraphrase per length of total lexicon. This is given by the zero temperature limit

$$D_{min} = \lim_{\beta \to \infty} f_{\beta}.$$
 (B6)

This can be converted easily to a more reasonable measure which is the distortion per dimension of the embedding by dividing the RHS by  $\alpha$ . Another possible measure is the distortion per length of original phrase, which requires just dividing the RHS by  $\ell = L/N$ . Another interesting quantity is the average energy (distortion)

$$D = \langle H \rangle = \partial_{\beta} \left(\beta f_{\beta}\right) \tag{B7}$$

### 1. Replica Partition function

Here we calculate the quenched averaging of the replicated partition function. First, the replicated partition function takes the form

$$Z^n_\beta(c,E) = \sum_{\bar{c}^a_i} \int d\bar{x}^a \exp\left[-\frac{\beta}{2N} \sum_a \sum_{\mu=1}^P (\sigma^a \cdot E^\mu)^2 + i\bar{x}^a \left(\sum_i \bar{c}^a_i - \bar{L}\right)\right]$$
(B8)

$$=\sum_{\bar{c}_i^a} \int \prod_a d\bar{x}^a \prod_\mu Du^a_\mu \exp\left[i\sqrt{\frac{\beta}{N}} \sum_{\mu,a} u^a_\mu (\sigma^a \cdot E^\mu) + i\bar{x}^a \left(\bar{L} - \sum_i \bar{c}^a_i\right)\right],\tag{B9}$$

where we have linearized the argument of the exponential using a Hubbard-Stratonovich transformation, and have introduced for notational shorthand the Gaussian measure  $Du = d^{-u^2/2}/\sqrt{2\pi}$  such that  $\int Du = 1$ . We assume the embeddings are drawn randomly i.i.d. from a non-centered Gaussian

$$\langle E_i^{\mu} \rangle = b^{\mu}, \quad \langle \left( E_i^{\mu} \right)^2 \rangle_c = \sigma^2.$$
 (B10)

After averaging the replica partition function over E, and defining our order parameter

$$q_{ab} = \frac{1}{N} \sum_{i} \sigma_i^a \sigma_i^b. \tag{B11}$$

We get the averaged partition function

$$\langle Z^n_{\beta}(c,E)\rangle_E = \sum_{\bar{c}^a_i} \int \prod_a d\bar{x}^a \prod_\mu Du^a_\mu \exp\left[i\sqrt{\frac{\beta}{N}} \sum_{\mu,a} u^a_\mu (\sum_i \sigma^a_i)b^\mu - \frac{\beta\sigma^2}{2} \sum_\mu \sum_{a,b} u^a_\mu u^b_\mu q_{ab} + i\bar{x}^a \left(\bar{L} - \sum_i \bar{c}^a_i\right)\right]. \tag{B12}$$

There are a few simplifications to be made at the present time. First, we use the fact that the constrains on the  $L_1$  norm of the c and  $\bar{c}$  mean that

$$\sum_{i} \sigma_{i}^{a} = ||c||_{1} - ||\bar{c}^{a}||_{1} = L - \bar{L} = Nm, \quad m \equiv \ell - \bar{\ell},$$
(B13)

where we have introduced a parameter m which measures the difference in magnetization density of each configuration. Integrating the  $u^a_{\mu}$  we end up with

$$\langle Z^n_{\beta}(c,E)\rangle_E = \sum_{\bar{c}^a_i} \int \prod_a d\bar{x}^a \exp\left[-\frac{P}{2}\log\det\left(1+\beta\sigma^2q\right) - \frac{N\beta m^2\left(\sum_{\mu} b^2_{\mu}\right)}{2} \mathbf{l}^T K \mathbf{l} + i\bar{x}^a \left(\bar{L} - \sum_i \bar{c}^a_i\right)\right],\tag{B14}$$

where  $K = (1 + \beta \sigma^2 q)^{-1}$ , and  $\mathbf{l} \in \mathbb{R}^n$  with  $\mathbf{l}_a = 1$ . Next, we seek to reduce the redundancy in model parameters. Defining

$$\mu = \frac{\sum_{\mu} b_{\mu}^2}{\sigma^2},\tag{B15}$$

and redefining  $\beta \sigma^2 \rightarrow \beta$ , we get

$$\langle Z^n_\beta(c,E)\rangle_E = \sum_{\bar{c}^a_i} \int \prod_a d\bar{x}^a \exp\left[-N\frac{\alpha}{2}\log\det\left(1+\beta q\right) - \frac{N\beta\mu m^2}{2}\mathbf{l}^T K\mathbf{l} + i\bar{x}^a \left(\bar{L} - \sum_i \bar{c}^a_i\right)\right]. \tag{B16}$$

In order to now average over configurations  $\bar{c}$ , we introduce the Lagrange multiplier  $\hat{q}$  to enforce the constraint B11. This will render the partition function

$$\langle Z_{\beta}^{n}(c,E)\rangle_{E} = \int d[x,\hat{q},q] \exp\left[iN\sum_{a\leq b}\hat{q}_{ab}q_{ab} - N\frac{\alpha}{2}\log\det\left(1+\beta q\right) - \frac{N\beta\mu m^{2}}{2}\mathbf{l}^{T}K\mathbf{l} + i\bar{L}\sum_{a}\bar{x}^{a} + \sum_{i}\log\mathcal{Z}[\hat{q},\bar{x},c_{i}]\right],\tag{B17}$$

where we have denoted the integration measure

$$d[x,\hat{q},q] = \prod_{a} dx^{a} \prod_{a \le b} d\hat{q}_{ab} dq_{ab}, \tag{B18}$$

and we have introduced

$$\mathcal{Z}[\hat{q}_{ab}, \bar{x}^a, c_i] = \sum_{\bar{c}^a} \exp\left(-i\sum_{a\leq b} \hat{q}_{ab}(c_i - \bar{c}^a)(c_i - \bar{c}^b) - i\bar{x}^a\bar{c}^a\right).$$
(B19)

Now note that since  $c_i$  can only take on two values, and this single-site partition function does not otherwise depend on the site index, it similarity will only take on two values:  $\mathcal{Z}[\hat{q}, \bar{x}, 0]$  and  $\mathcal{Z}[\hat{q}, \bar{x}, 1]$ . Furthermore, since the norm of c is imposed to be L, we have that

$$\prod_{i} \mathcal{Z}[\hat{q}_{ab}, \bar{x}^{a}, c_{i}] = \left(\mathcal{Z}[\hat{q}_{ab}, \bar{x}^{a}, 1]\right)^{L} \left(\mathcal{Z}[\hat{q}_{ab}, \bar{x}^{a}, 0]\right)^{N-L}.$$
(B20)

Therefore, the partition function does not depend on the detailed configuration of c, but only on its total length L. This means we can trivially take the average over c and get

$$\langle Z^n_{\beta}(c,E)\rangle_{c,E} = \int d[x,\hat{q},q] \exp\left[iN\sum_{a\leq b}\hat{q}_{ab}q_{ab} - N\frac{\alpha}{2}\log\det\left(1+\beta q\right) - \frac{N\beta\mu m^2}{2}\mathbf{l}^T K\mathbf{l} + i\bar{L}\sum_a \bar{x}^a \tag{B21}\right]$$

$$+L\log \mathcal{Z}[\hat{q}, \bar{x}, 1] + (N - L)\log \mathcal{Z}[\hat{q}, \bar{x}, 0] \Big].$$
(B22)

Next, we find explicit expressions for  $\mathcal{Z}$  that we can work with. Defining

$$\mathcal{Z}[\hat{q},\bar{x}] \equiv \mathcal{Z}[\hat{q}_{ab},\bar{x}^a,0] = \sum_{\bar{c}^a} \exp\left(-i\sum_{a\le b} \hat{q}_{ab}\bar{c}^a\bar{c}^b - i\sum_a \bar{x}^a\bar{c}^a\right),\tag{B23}$$

we find that

$$\mathcal{Z}[\hat{q}_{ab}, \bar{x}^a, 1] = e^{-i\sum_a \bar{x}^a} \mathcal{Z}[\hat{q}, -\bar{x}].$$
(B24)

(B25)

With this, we get the replica partition function in a form which will allow us to perform a saddle-point calculation

$$\langle Z_{\beta}^{n}(c,E)\rangle_{c,E} = \int d[x,\hat{q},q] \exp\left(N\mathcal{H}_{n}\right), \tag{B26}$$

where

$$\mathcal{H}_n = -\frac{\alpha}{2}\log\det\left(1+\beta q\right) - \frac{\beta\mu m^2}{2}\ell^T K\ell + \sum_{a\leq b}Q_{ab}q_{ab} - \sum_a X^a m + t\log\mathcal{Z}[Q, -X] + (1-t)\log\mathcal{Z}[Q, X], \tag{B27}$$

$$K = (1 + \beta q)^{-1}, \quad \ell_i = 1, \quad m = \ell - \overline{\ell}, \quad Q \equiv i\hat{q}, \quad X \equiv i\overline{x}, \tag{B28}$$

and the "single site partition function" is

$$\mathcal{Z}[Q,X] = \sum_{s^a \in \{0,1\}} e^{-\sum_{a \le b} Q_{ab} s^a s^b - \sum_a X^a s^a}.$$
(B29)

Now defining for some function of spins G(s), the single site correlation function is defined as

$$\langle G(s) \rangle_{\pm} \equiv \frac{1}{\mathcal{Z}[Q, \pm X]} \sum_{s^a \in \{0, 1\}} G(s) e^{-\sum_{a \le b} Q_{ab} s^a s^b - \sum_a (\pm X^a) s^a}.$$
 (B30)

The saddle-point equations are

$$\frac{\delta}{\delta Q_{ab}} \mathcal{H}_n = q_{ab} - \ell \langle s^a s^b \rangle_- - (1 - \ell) \langle s^a s^b \rangle_+, \tag{B31}$$

$$\frac{\delta}{\delta X_a} \mathcal{H}_n = -m + \ell \langle s^a \rangle_- - (1 - \ell) \langle s^a \rangle_+, \tag{B32}$$

$$\frac{\delta}{\delta q_{ab}} \mathcal{H}_n = -\alpha \beta K_{ab} + \beta^2 \mu m^2 (K1)_a (K1)_b + Q_{ab}, \quad a \neq b,$$
(B33)

$$\frac{\delta}{\delta q_{aa}} \mathcal{H}_n = -\frac{\alpha\beta}{2} K_{ab} + \frac{\beta^2 \mu m^2}{2} (K\mathbf{1})_a (K\mathbf{1})_b + Q_{aa}.$$
(B34)

### 2. Replica symmetric ansatz

Assuming replica symmetric order parameters

$$q_{ab} = (q_0 - q)\delta_{ab} + q, \quad Q_{ab} = (Q_0 - Q)\delta_{ab} + Q, \quad X^a = X,$$
(B35)

implies also that

$$K_{ab} = (K_0 - K)\delta_{ab} + K. \tag{B36}$$

We label  $\Delta = q_0 - q$ , since this appears very often below. Then the components of the inverse propagator K are

$$K_0 - K = \frac{1}{1 + \beta \Delta}, \quad K = -\frac{\beta q}{(1 + \beta \Delta)(1 + \beta \Delta + n\beta q)}.$$
 (B37)

We first evaluate the single-site partition function

$$\mathcal{Z}[Q,X] = \sum_{s^a} \exp\left[-(Q_0 - \frac{1}{2}Q + X)\sum_a s_a - \frac{1}{2}Q\left(\sum_a s_a\right)^2\right],$$
(B38)

$$= \int Dz \sum_{s^{a}} e^{\sqrt{-Q}z \sum_{a} s_{a} - (Q_{0} - \frac{1}{2}Q + X) \sum_{a} s_{a}}, \tag{B39}$$

$$= \int Dz \left(1 + e^{-\Theta_{+}(z)}\right)^{n}, \quad \Theta_{\pm}(z) = -\sqrt{-Q}z + Q_{0} - \frac{1}{2}Q \pm X, \tag{B40}$$

$$\rightarrow 1 + n \int Dz \log\left(1 + e^{-\Theta_+(z)}\right) + O(n^2). \tag{B41}$$

On the replica symmetric saddle,  $\mathcal{H}_n^* = n\mathcal{H}_0^* + O(n^2)$ , where

$$\mathcal{H}_{0}^{*} = -\frac{\alpha}{2} \left[ \log(1+\beta\Delta) + \frac{\beta q}{1+\beta\Delta} \right] - \frac{\beta\mu m^{2}}{2} \frac{1}{1+\beta\Delta} + Q_{0}q_{0} - \frac{1}{2}Qq - Xm$$
(B42)

$$+\ell \int Dz \log(1+e^{-\Theta_{-}(z)}) + (1-\ell) \int Dz \log(1+e^{-\Theta_{+}(z)}) + O(n^{2}),$$
(B43)

so that the free energy density becomes a function of the relative message and compression lengths,  $\ell$  and  $\bar{\ell}$ , respectively, as well as the relative embedding dimension  $\alpha = P/N$ , and the average mean of the embedding vectors  $\mu$ :

$$f_{\beta}(\ell,\bar{\ell},\alpha,\mu) = -\frac{1}{\beta}\mathcal{H}_{0}^{*} = \frac{\alpha}{2\beta} \left[ \log(1+\beta\Delta) + \frac{\beta q}{1+\beta\Delta} \right] + \frac{\mu m^{2}}{2} \frac{1}{1+\beta\Delta} - \frac{1}{\beta}Q_{0}q_{0} + \frac{1}{2\beta}Qq + \frac{1}{\beta}Xm$$
(B44)

$$-\frac{1}{\beta}\ell\int Dz\log(1+e^{-\Theta_{-}(z)}) - \frac{1}{\beta}(1-\ell)\int Dz\log(1+e^{-\Theta_{+}(z)}) + O(n^{2}).$$
(B45)

The saddle-point equations become

$$Q_0 = \frac{\alpha\beta}{2} \left[ \frac{1}{1+\beta\Delta} - \frac{\beta q}{(1+\beta\Delta)^2} \right] - \frac{\beta^2 \mu m^2}{2(1+\beta\Delta)^2},\tag{B46}$$

$$-Q = \frac{\beta^2 (\alpha q + \mu m^2)}{(1 + \beta \Delta)^2},\tag{B47}$$

$$\ell - \bar{\ell} = \ell \int Dz \frac{1}{1 + e^{\Theta_{-}(z)}} - (1 - \ell) \int Dz \frac{1}{1 + e^{\Theta_{+}(z)}}, \tag{B48}$$

$$q_0 = \ell \int Dz \frac{1}{1 + e^{\Theta_-(z)}} + (1 - \ell) \int Dz \frac{1}{1 + e^{\Theta_+(z)}},$$
(B49)

$$q = \ell \int Dz \frac{1}{\left(1 + e^{\Theta_{-}(z)}\right)^{2}} + (1 - \ell) \int Dz \frac{1}{\left(1 + e^{\Theta_{+}(z)}\right)^{2}}.$$
 (B50)

Plugging these into the free energy density affords some simplifications:

$$f_{\beta}(\ell,\bar{\ell},\alpha,\mu) = \frac{\alpha}{2\beta}\log(1+\beta\Delta) - \frac{\alpha\Delta}{2(1+\beta\Delta)^2}\left(1+\beta\Delta-\beta q\right) + \frac{\mu m^2}{2}\frac{(1+2\beta\Delta)}{(1+\beta\Delta)^2} \tag{B51}$$

$$+\frac{X(\ell-\bar{\ell})}{\beta} - \frac{\ell}{\beta} \int Dz \log\left(1 + e^{-\Theta_{-}(z)}\right) - \frac{(1-\ell)}{\beta} \int Dz \log\left(1 + e^{-\Theta_{+}(z)}\right). \tag{B52}$$

Summary of Replica Symmetric Mean Field Theory: The self-consistent mean-field equations are

$$\ell - \bar{\ell} = \ell \int Dz \frac{1}{1 + e^{\Theta_{-}(z)}} - (1 - \ell) \int Dz \frac{1}{1 + e^{\Theta_{+}(z)}}, \tag{B53}$$

$$q_0 = \ell \int Dz \frac{1}{1 + e^{\Theta_-(z)}} + (1 - \ell) \int Dz \frac{1}{1 + e^{\Theta_+(z)}},$$
(B54)

$$q = \ell \int Dz \frac{1}{\left(1 + e^{\Theta_{-}(z)}\right)^{2}} + (1 - \ell) \int Dz \frac{1}{\left(1 + e^{\Theta_{+}(z)}\right)^{2}}.$$
 (B55)

where

$$\Theta_{\pm}(z) = -\frac{\beta}{1+\beta\Delta}\sqrt{\alpha q + \mu m^2}z + \frac{\alpha\beta}{2(1+\beta\Delta)} \pm X,$$
(B56)

and the free energy density is

$$f = \frac{\alpha}{2\beta} \log(1 + \beta\Delta) - \frac{\alpha\Delta}{2(1 + \beta\Delta)^2} \left(1 + \beta\Delta - \beta q\right) + \frac{\mu m^2}{2} \frac{(1 + 2\beta\Delta)}{(1 + \beta\Delta)^2}$$
(B57)

$$+\frac{X(\ell-\bar{\ell})}{\beta} - \frac{\ell}{\beta} \int Dz \log\left(1 + e^{-\Theta_{-}(z)}\right) - \frac{(1-\ell)}{\beta} \int Dz \log\left(1 + e^{-\Theta_{+}(z)}\right).$$
(B58)

Differentiating the free energy density also gives the average distortion quoted in the main text:

$$D(T) = \frac{1}{2(1+\beta\Delta)^2} \left[ \alpha(q_0 + \beta\Delta^2) + \mu m^2 \right].$$
(B59)

# 3. Zero temperature limits

In the zero temperature limit, we assume

$$q_0 - q = T\delta + O(T^2), \tag{B60}$$

and we change variables

$$X \equiv \frac{\beta\lambda}{1+\beta\Delta},\tag{B61}$$

which allows us to write

$$\Theta_{\pm}(z) = \frac{\beta}{1+\beta\Delta} \left[ -\sqrt{\alpha q + \mu m^2} z + \frac{\alpha}{2} \pm \lambda \right] = \frac{1}{T} \left[ -A_0 z + B(\pm\lambda) \right] - A_1 z + O(T^2), \tag{B62}$$

$$A_0 = \frac{\sqrt{\alpha \bar{q} + \mu m^2}}{1 + \delta}, \quad B(\lambda) = \frac{\alpha/2 + \lambda}{1 + \delta}.$$
(B63)

Note that without the hard constraint on the compression length, we would not be permitted to make such a change. This change of variables ends up being very useful in finding a simple expression for  $\delta$ . Before we find this expression, we first show that  $\Delta = O(T)$ . Using the MFT to write  $\Delta$ ,

$$\Delta = q_0 - q = \ell \int Dz \frac{e^{\Theta_{-}(z)}}{(1 + e^{\Theta_{-}(z)})^2} + (1 - \ell) \int Dz \frac{e^{\Theta_{+}(z)}}{(1 + e^{\Theta_{+}(z)})^2}.$$
 (B64)

Now changing variables  $z = (B(\lambda) - T\theta)/A_0$ , and neglecting the extra T dependence in  $\Theta$ , we get

$$\int Dz \frac{e^{\Theta(z)}}{(1+e^{\Theta(z)})^2} = \int \frac{1}{\sqrt{2\pi}} \frac{Td\theta}{A_0} e^{-\frac{1}{2A_0^2}(T\theta - B(\lambda))^2} \frac{e^{\theta}}{(1+e^{\theta})^2} = T \frac{1}{\sqrt{2\pi}A_0} e^{-\frac{B(\lambda)^2}{2A_0^2}} \int d\theta \left(-\partial_{\theta}\right) \frac{1}{1+e^{\theta}} + O(T^2) \quad (B65)$$

$$=T\frac{1}{\sqrt{2\pi}A_0}e^{-\frac{B(\lambda)^2}{2A_0^2}} + O(T^2)$$
(B66)

Expanding  $\Delta$  to order T then gives

$$\delta = \frac{\ell}{\sqrt{2\pi}A_0} e^{-\frac{B(-\lambda)^2}{2A_0^2}} + \frac{(1-\ell)}{\sqrt{2\pi}A_0} e^{-\frac{B(\lambda)^2}{2A_0^2}} = \ell(1+\delta)H_2(-\lambda) + (1-\ell)(1+\delta)H_2(\lambda), \tag{B67}$$

where

$$H_2(\lambda) \equiv \frac{1}{\sqrt{2\pi(\alpha\bar{q} + \mu m^2)}} \exp\left(-\frac{(\alpha/2 + \lambda)^2}{2(\alpha\bar{q} + \mu m^2)}\right)$$
(B68)

This is a linear equation for  $\delta$  that can be solved to give:

$$\delta = \frac{\ell H_2(-\lambda) + (1-\ell)H_2(\lambda)}{1 - \ell H_2(-\lambda) - (1-\ell)H_2(\lambda)}.$$
(B69)

(B70)

We also have in the zero temperature limit

$$H_1(\lambda) = \int \frac{1}{1 + e^{\Theta(z,\lambda)}} Dz = \int_{B(\lambda)/A_0} Dz = \frac{1}{2} \operatorname{erfc}\left(\frac{B(\lambda)}{\sqrt{2}A_0}\right),$$
(B71)

(B72)

so that the zero temperature order parameter is

$$\bar{q} = \ell H_1(-\lambda) + (1-\ell)H_1(\lambda), \tag{B73}$$

$$\ell - \bar{\ell} = \ell H_1(-\lambda) - (1-\ell)H_1(\lambda). \tag{B74}$$

(B75)

Finally, we use the following small T expansion:

$$\int Dz \log\left(1 + e^{-\Theta(z,\lambda)}\right) = \frac{A_0}{\sqrt{2\pi}} e^{-B(\lambda)^2/2A_0} - B(\lambda) \frac{1}{2} \operatorname{erfc}(B(\lambda)/\sqrt{2}A_0), \tag{B76}$$

$$=\frac{1}{T(1+\delta)}\left[\left(\alpha\bar{q}+\mu m^2\right)J_2(\lambda)-\left(\frac{\alpha}{2}+\lambda\right)J_1(\lambda)\right]+O(1),\tag{B77}$$

to write the zero temperature free energy density

$$\lim_{T \to 0} f = \frac{\alpha \bar{q}\delta}{2(1+\delta)^2} + \frac{\mu m^2}{2} \frac{(1+2\delta)}{(1+\delta)^2} + \frac{\lambda(\ell-\bar{\ell})}{1+\delta} - \frac{\ell}{(1+\delta)} \left[ (\alpha \bar{q} + \mu m^2) H_2(-\lambda) - (\frac{\alpha}{2} - \lambda) H_1(-\lambda) \right]$$
(B78)

$$-\frac{(1-\ell)}{(1+\delta)}\left[\left(\alpha\bar{q}+\mu m^2\right)H_2(\lambda)-\left(\frac{\alpha}{2}+\lambda\right)H_1(\lambda)\right].$$
 (B79)

Summary of zero temperature MFT in lossy phase: At zero temperature, we have

$$\bar{q} = \ell H_1(-\lambda) + (1-\ell)H_1(\lambda), \tag{B80}$$

$$\ell - \bar{\ell} = \ell H_1(-\lambda) - (1 - \ell) H_1(\lambda) \tag{B81}$$

$$\delta = \frac{\ell H_2(-\lambda) + (1-\ell)H_2(\lambda)}{1-\ell H_2(-\lambda) - (1-\ell)H_2(\lambda)}$$
(B82)

with

$$H_1(\lambda) = \frac{1}{2} \operatorname{erfc}\left[\frac{\alpha/2 + \lambda}{\sqrt{2(\alpha\bar{q} + \mu m^2)}}\right], \quad H_2(\lambda) = \frac{1}{\sqrt{2\pi(\alpha\bar{q} + \mu m^2)}} \exp\left(-\frac{(\alpha/2 + \lambda)^2}{2(\alpha\bar{q} + \mu m^2)}\right)$$
(B83)

Solving zero temperature MFT: Here we provide some details on how we find solutions for the MFT. We focus on the setting with  $\mu = 0$ , and in the lossy limit. Our goal is to simplify the MFT equations in this phase Eqs. B80, B81,B82. First, define

$$x \equiv H_1(-\lambda) = \frac{1}{2} \operatorname{erfc}\left(\frac{\alpha/2 - \lambda}{\sqrt{2\alpha\bar{q}}}\right), \quad y \equiv H_1(\lambda).$$
(B84)

From these, the constraint equation gives

$$y(x) = \frac{1}{1-\ell} \left( \ell x - \ell + \bar{\ell} \right) = \frac{\ell}{1-\ell} \left( x - 1 + C \right),$$
(B85)

and the order parameter becomes

$$\bar{q} = \ell x + (1 - \ell) y(x) = \ell (2x + C - 1).$$
(B86)

Next, we use

$$\frac{\alpha/2 - \lambda}{\sqrt{2\alpha\bar{q}}} = \operatorname{erfc}^{-1}(2x), \quad \frac{\alpha/2 + \lambda}{\sqrt{2\alpha\bar{q}}} = \operatorname{erfc}^{-1}(2y(x)).$$
(B87)

It is important that we select the branch which has

$$\operatorname{erfc}^{-1}(2x) + \operatorname{erfc}^{-1}(2y(x)) > 0.$$
 (B88)

which implies a constraint

$$x < 1 - \bar{\ell}.\tag{B89}$$

In addition, we require y(x) > 0, which requires

$$x > 1 - C. \tag{B90}$$

from Eq. B87, we get

$$\bar{q} = \frac{\alpha}{2} \left( \operatorname{erfc}^{-1}(2x) + \operatorname{erfc}^{-1}(2y(x)) \right)^{-2},$$
(B91)

combining this with Eq. B86 gives the implicit equation for  $x \in [1 - C, 1 - \overline{\ell}]$ :

$$(2x + C - 1) = \frac{\alpha}{2\ell} \left( \operatorname{erfc}^{-1}(2x) + \operatorname{erfc}^{-1}(2y(x)) \right)^{-2}.$$
 (B92)

We see from this that solutions to x depend on  $\alpha$  only through the ratio  $\alpha/\ell$ . However, they do depend on the absolute value of  $\ell$  as well, through the ratio  $\ell/(1-\ell)$ .

Finally, an additional condition for a self-consistent solution is that  $Q_1 > 0$ , This turns into the following inequality:

$$1 - \ell \frac{1}{\sqrt{2\pi\alpha\bar{q}}} e^{-(\text{erfc}^{-1}(2x))^2} - (1 - \ell) \frac{1}{\sqrt{2\pi\alpha\bar{q}}} e^{-(\text{erfc}^{-1}(2y(x)))^2} > 0.$$
(B93)

The critical curve is given by the implicit equation:

$$\sqrt{2\pi\alpha(2x+C-1)/\ell} = e^{-\left(\operatorname{erfc}^{-1}(2x)\right)^2} + \frac{(1-\ell)}{\ell} e^{-\left(\operatorname{erfc}^{-1}(2y(x))\right)^2}.$$
(B94)

We compute these curves numerically for different  $\ell$  on the  $\alpha$ -C plane in Fig. 3, and plot them with solid colored lines. There is another transition for compression ratios closer to unity, indicated by dashed lines in Fig. 3, in which there is a bifurcation in the solutions to B92. This is therefore a topological transition: below the dashed curves, there is only one solution to B92, whereas above it there are three. The point where the solid curve hits the dashed curve (shown with a star in the figure) represents the maximal value of  $\alpha$  for which lossless compression can occur. In the right panel of Fig. 3, we show that this maximum value depends on the total message length  $\ell$ , but has an upper bound and never exceeds  $\alpha \approx 0.4049$  for all  $\ell$ . We argue that this value has to occur at  $\lambda = 0$ . In this case, we get  $\operatorname{erfc}^{-1} y(x) = \operatorname{erfc}^{-1} x$ ,  $\bar{q} = x$ , C = 1 - x, and we must solve both of the following equations simultaneously to find this maximal  $\alpha$ :

$$\sqrt{2\pi\alpha x} = e^{-\left(\operatorname{erfc}^{-1}(2x)\right)^2}, \quad x = \frac{\alpha}{4\left(\operatorname{erfc}^{-1}(2x)\right)^2}.$$
 (B95)

### 4. Zero Compression Limit: C = 1

a. Half Filling The easiest setting to consider is half-filling, i.e.  $\ell = 1/2$ . In this case, with C = 1, the constraint equation implies that  $\lambda = 0$ , and that

$$\bar{q} = \frac{1}{2} \operatorname{erfc}\left(\frac{\alpha}{2\sqrt{2\alpha\bar{q}}}\right) \tag{B96}$$

This has nonzero solutions up to  $\alpha \approx 0.6629$ . Above this, the only solution is  $\bar{q} = 0$ , which also has zero entropy. This corresponds to zero distortion. In the compressed sensing problem, this is akin to perfect reconstruction. Below this limit, the finite entropy solution for  $\bar{q}$  has nonzero distortion.

b. Arbitrary Filling Now consider arbitrary  $\ell$ . This requires keeping  $\lambda$ . However, some simplifications occur. For instance,

$$\ell H_1(-\lambda) = (1-\ell)H_1(\lambda) \tag{B97}$$

which implies

$$\bar{q} = 2\ell H_1(-\lambda) \tag{B98}$$

In general, the transition to perfect reconstruction (above which the only solution is  $\bar{q} = 0$ ) occurs at

$$\alpha/\ell \approx 1.3257\tag{B99}$$

For larger values of  $\alpha$ ,  $\bar{q} = 0$ . This means that for P/L > 1.35, the typical configuration recovers the original message exactly. This occurs when the dimensionality of the embedding space is larger than the message length, suggesting that it is a limit in which the message vectors are statistically orthogonal. In the opposite limit, P < L, the message must utilize many linearly dependent vectors to construct the message.



FIG. 3: Phase diagram dependence on message length (Left) The zero temperature phase diagram of the RS MFT for different values of relative message length  $\ell$ . In each figure, we plot only the curves demarcating the compressible phase Q > 0. The solid curve indicates the discontinuous transition, whereas the dashed curve corresponds the appearance of self-consistent incompressible (Q = 0) solutions. For each fixed  $\ell$ , the compressible region extends out to some maximal  $\alpha$  corresponding to the point where these two curves meet (denoted by a star in the figure). (**Right**) shows that the maximal  $\alpha$  depends non-monotonically on  $\ell$ , and tends to zero both as  $\ell \to 0$  and  $\ell \to 1$ .