An Indirect Rate-Distortion Characterization for Semantic Sources: General Model and the Case of Gaussian Observation

Jiakun Liu, Shuo Shao, *Member, IEEE*, Wenyi Zhang, *Senior Member, IEEE*, and H. Vincent Poor, *Life Fellow, IEEE*

***Abstract*— A new source model, which consists of an intrinsic state part and an extrinsic observation part, is proposed and its information-theoretic characterization, namely its rate-distortion function, is defined and analyzed. Such a source model is motivated by the recent surge of interest in the semantic aspect of information: the intrinsic state corresponds to the semantic feature of the source, which in general is not observable but can only be inferred from the extrinsic observation. There are two distortion measures, one between the intrinsic state and its reproduction, and the other between the extrinsic observation and its reproduction. Under a given code rate, the tradeoff between these two distortion measures is characterized by the rate-distortion function, which is solved via the indirect rate-distortion theory and is termed the semantic rate-distortion function of the source. As an application of the general model and its analysis, the case of Gaussian extrinsic observation is studied, assuming a linear relationship between the intrinsic state and the extrinsic observation, under a quadratic distortion structure. The semantic rate-distortion function is shown to be the solution of a convex programming problem with respect to an error covariance matrix, and a reverse water-filling type of solution is provided when the model further satisfies a diagonalizability condition.**

***Index Terms*— Lossy compression, rate distortion theory, reverse water-filling, semantic rate distortion function, semantic source model, task-oriented communication.**

Manuscript received 28 January 2022; revised 1 June 2022; accepted 25 July 2022. Date of publication 29 July 2022; date of current version 16 September 2022. The work of Jiakun Liu and Wenyi Zhang was supported in part by the National Key Research and Development Program of China under Grant 2018YFA0701603; the work of Shuo Shao was supported in part by the National Key Research and Development Program of China under Grant 2020YFA0712302, and the National Natural Science Foundation of China under Grant 61901261; and Natural Science Foundation of Shanghai under Grant 19YF1424200; and the work of H. Vincent Poor was supported in part by the U.S. National Science Foundation under Grant CNS-2128448. An earlier version of this paper was presented in part at the IEEE International Symposium on Information Theory (ISIT), 2021 [DOI: 10.1109/ISIT45174.2021.9518240]. The associate editor coordinating the review of this article and approving it for publication was Y. Fang. *(Jiakun Liu and Shuo Shao are co-first authors.) (Corresponding authors: Wenyi Zhang; H. Vincent Poor.)*

Jiakun Liu and Wenyi Zhang are with the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027, China (e-mail: liujk@mail.ustc.edu.cn; wenyizha@ustc.edu.cn).

Shuo Shao is with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: shuoshao@sjtu.edu.cn).

H. Vincent Poor is with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: poor@princeton.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCOMM.2022.3194978.

Digital Object Identifier 10.1109/TCOMM.2022.3194978

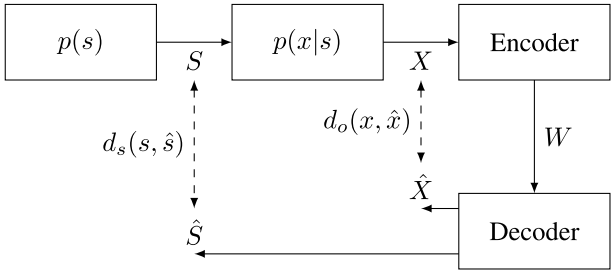


Fig. 1. Illustration of a semantic source and its lossy compression.

I. Introduction

A STANDARD approach to describe an information source is to model a source as a stochastic process $\{X_i\}$, and when the stochastic process is memoryless, it suffices to model a source as a random variable[1] X with a given probability distribution $p(x)$ [2], [3]. In this paper, we study a new source model, which consists of an intrinsic state process and an extrinsic observation process. In the memoryless case, we can describe such a source model as a pair of random variables (S, X), with a given joint probability distribution $p(s, x)$, defined over an appropriate product alphabet $\mathcal{S} \times \mathcal{X}$.

In order to characterize the information-theoretic aspect of such a source, consider the problem of compressing the source (S, X) so as to reproduce, in a lossy sense, a reproduction $(\hat{S}, \hat{X})$ over a reproduction product alphabet $\hat{\mathcal{S}} \times \hat{\mathcal{X}}$. Of course, a pair of distortion measures, $d_s : \mathcal{S} \times \hat{\mathcal{S}} \mapsto \mathbb{R}$ and $d_o : \mathcal{X} \times \hat{\mathcal{X}} \mapsto \mathbb{R}$, are introduced correspondingly. Here, the subscript s stands for "state" and the subscript o stands for "observation". A key point of the problem is that the compressor only has access to X, the extrinsic observation; — while S, the intrinsic state, remains unrevealed. The situation is illustrated in Figure 1.

Our source model, termed a semantic source in the sequel, is motivated by the recent surge of interest in the semantic aspect of information. In a number of applications that may benefit from taking into account the "semantic" feature of information, it is adequate to adopt a goal-oriented perspective; that is, the destination's interest in obtaining a piece of information is to accomplish a certain goal. Furthermore, it is customary to adopt an inference-theoretic problem formulation,

[1] In this paper, random variables can be drawn from general alphabets, so random vectors are vector-valued random variables.

which casts the accomplishment of the said goal as solving a statistical inference problem. The reproduction of the intrinsic state S corresponds to the semantic inference part of the source, and the reproduction of the extrinsic observation X corresponds to the conventional lossy compression part of the source.

We give two examples of the above consideration:

- Systems that support MPEG Video Coding for Machines (VCM) are becoming popular in applications. In VCM, both the video itself and its features are reproduced: the video signal is for human vision, and the features are for machine vision tasks [4], [5] [6]. Treating the video as a semantic source, the video signal itself corresponds to its extrinsic observation, and the underlying features correspond to its intrinsic state, so as to embody the semantic aspect of the video. Usually the code rate required for reproducing features can be drastically lower than that required for reproducing the video signal itself. Intuitively, features typically have much smaller rate distortion functions and hence can be described with many fewer bits, compared with video signals. For instance, previous works have shown that neural network-based learning techniques can extract a very small amount of data from video signals to satisfy the need of action recognition, target classification and many other tasks [7], [8]. In contrast, traditional video coding schemes such as H.264/AVC/MPEG-4 and H.265/HEVC/MPEG-H Part 2 target reproducing the video signal with high fidelity, but may perform poorly for machine vision purposes [9].
- In coding of speech signals, the semantic aspect is embodied as a sequence of text words, which, of course, can only be inferred from the speech signal itself. Treating the speech as a semantic source, the words correspond to its intrinsic state and the speech signal corresponds to its extrinsic observation. It is the usual case that both the words and the speech signal are desirable, because the words carry the meaning of speech, and the speech signal waveform may help us infer the stress and emotion of the speaker [10], and may further help us accomplish tasks like speaker recognition and speaker verification [11].

Our main contributions include:

- We propose a theoretical framework based on rate distortion theory for characterizing semantic information.
- We define and derive a single-letter expression for the semantic rate distortion function.
- When the extrinsic observation is Gaussian and satisfies a linear relationship with the intrinsic state, we reduce the calculation of the semantic rate distortion function to a convex programming problem, which is tractable with standard scientific computing software. Furthermore, under a diagonalizability condition, we obtain a weighted reverse water-filling solution for the semantic rate distortion function.

We give a brief overview of related works in the remaining part of this section. Then we provide a formal mathematical description of the semantic source model and the corresponding semantic rate-distortion problem formulation in Section II, for which we establish the semantic rate-distortion function in general form in Section III. As an application of the general results, in Section IV we turn to a case study of Gaussian extrinsic observation, assuming a linear relationship between the intrinsic state and the extrinsic observation, under a quadratic distortion structure. Therein, we formulate a convex programming problem to solve for the semantic rate-distortion function. When the Gaussian observation model further satisfies a diagonalizability condition, we develop a reverse water-filling type of solution in Section V. Finally we conclude this paper in Section VI.

A. Related Works

The first formulation in Shannon's information theory is lossless source coding, wherein a sequence of symbols obeying a certain probabilistic law is represented as a bit string (i.e., a codeword) by an encoder, and the decoder reproduces, based upon the codeword, the original sequence of symbols, with success probability exactly one or asymptotically approaching one. Hence, the coding is solely determined by the probabilistic model of the source, and there is certainly no role of the semantic aspect of the source. This is also consistent with Shannon's remark in his landmark paper [2], saying "these semantic aspects of communication are irrelevant to the engineering problem."

In a broad sense, however, the lossy source coding formulation in Shannon's information theory, namely, the rate-distortion theory [12], has provided a means of studying the semantic aspects of a source. This is because the coding is not solely determined by the probabilistic model of a source, but is also affected by a distortion measure, which may be defined in a rather versatile way so as to capture the "utility" when the source is reproduced at the decoder.

Our present work goes one step further, by endowing a source with a state-observation structure and studying the rate distortion function of such a source model. This model captures the fact that the semantic aspects of a source are generally embedded as intrinsic features, and hence should be characterized by studying the reproduction of the intrinsic state, in addition to the reproduction of the extrinsic observation. Our treatment of semantic aspects of sources is also in line with the recent heightened interest in the development of 5G and beyond wireless systems [13], [14] [15], where for many applications the semantic aspects correspond to the accomplishment of certain inference goals. Hence, if we consider an information theoretic characterization of such a "semantic" source, the task of coding is to efficiently encode the extrinsic observation so that the decoder can infer both the intrinsic state and the extrinsic observation, subject to fidelity criteria on both, simultaneously. Our problem formulation and approach are closely related to two variants of the standard rate distortion theory, namely, the indirect rate distortion function and the rate distortion function under multiple distortion measures; see our discussion following Theorem 1 in Section III.

The inference-theoretic goal-oriented approach adopted in our problem formulation does not seek a task-independent universal definition of semantic information, which is outside

the scope of the present paper; for some attempts in that regard, see, e.g., [16], [17] [18], [19] for a few representative works that undertake drastically different approaches.

As related topics, the information bottleneck [20], [21] and the privacy funnel [22], [23] are, in a certain sense, dual concepts, and both place constraints in terms of mutual information. The underlying idea of the information bottleneck is, in a broad sense, similar to ours. Specifically, there one generates a reproduction based upon the extrinsic observation, minimizing the mutual information between the extrinsic observation and the reproduction, while maintaining a level of mutual information between the intrinsic state and the reproduction. But for the information bottleneck problem formulation, there is neither an explicit distortion measure, nor an operational definition of lossy compression.

Task-based compression has been approached mainly from the perspective of quantizer design [24]. It has been demonstrated that steering the design goal according to the task leads to performance benefits compared with a conventional task-agnostic approach, a conclusion in line with what we advocate in our work. The perception-distortion tradeoff [25] imposes an additional constraint on the probability distribution of the reproduction. None of these related works proposes to decompose the information source into intrinsic and extrinsic parts as in our work, let alone investigates the joint behavior of them. In [26], a similar intrinsic state-extrinsic observation model is studied, but the encoder is designed based on the marginal distribution of the extrinsic observation only.

II. System Model and Problem Formulation

As already outlined in the introduction, we model a memoryless semantic source as a pair of random variables (S, X) that are correlated with joint probability distribution $p(s, x)$. The semantic aspect is embodied in the intrinsic state S, which is not observable but can only be inferred from the extrinsic observation X. In order to characterize the rate-distortion behavior of the semantic source, we consider a sequence of independent and identically distributed (i.i.d.) samples of (S, X), denoted as $(S_i, X_i)_{i\in\mathbb{N}}$, and denote its length-$n$ block as (S^n, X^n).

The i.i.d. source model is an idealistic scenario for our information-theoretic study. Real-world data generally exhibit sophisticated memory structures. A particularly interesting scenario is when the intrinsic state is a Markov chain, and the extrinsic observation obeys a hidden Markov model (HMM) [27]. Extensions of our approach for semantic source models with memory are left for future research.

The lossy compression of a semantic source has been illustrated in Figure 1. The encoder only has access to a length-n block of the extrinsic observation sequence X^n, and the decoder has two tasks: reproducing the intrinsic state block as $\hat{S}^n$ under a state distortion measure d_s, and reproducing the extrinsic observation block as $\hat{X}^n$ under an observation distortion measure d_o. The encoder and the decoder are connected via a bit pipe in which the codeword W of nR bits is transferred from the encoder to the decoder, where R is thus the code rate of the lossy compression system.

Below we provide a formal description of the lossy compression problem of a semantic source.

Let $d_s : \mathcal{S} \times \hat{\mathcal{S}} \to \mathbb{R}_+$ and $d_o : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}_+$ be two given distortion measures, defined over the source product alphabet $\mathcal{S} \times \mathcal{X}$ and the reproduction product alphabet $\hat{\mathcal{S}} \times \hat{\mathcal{X}}$. The extended block-wise distortion measures are as follows:

$$d_s(s^n, \hat{s}^n) = \frac{1}{n}\sum_{i=1}^{n} d_s(s_i, \hat{s}_i), \tag{1}$$

$$d_o(x^n, \hat{x}^n) = \frac{1}{n}\sum_{i=1}^{n} d_o(x_i, \hat{x}_i). \tag{2}$$

We claim a tuple (R, D_s, D_o) to be achievable, if for any $\epsilon > 0$ and all sufficiently large n, there exist the following functions:

- Encoding function $f : \mathcal{X}^n \mapsto \{1, 2, \ldots, 2^{\lfloor n(R+\epsilon)\rfloor}\}$ which generates the codeword W as $W = f(X^n)$;
- State decoding function $g_s : \{1, 2, \ldots, 2^{\lfloor n(R+\epsilon)\rfloor}\} \mapsto \hat{\mathcal{S}}^n$, such that
$$\mathbb{E}\left[d_s(S^n, \hat{S}^n)\right] \le D_s + \epsilon, \tag{3}$$
where $\hat{S}^n = g_s(f(X^n))$;
- Observation decoding function $g_o : \{1, 2, \ldots, 2^{\lfloor n(R+\epsilon)\rfloor}\} \mapsto \hat{\mathcal{X}}^n$, such that
$$\mathbb{E}\left[d_o(X^n, \hat{X}^n)\right] \le D_o + \epsilon, \tag{4}$$
where $\hat{X}^n = g_o(f(X^n))$.

It is clear that the state decoding function g_s and the observation decoding function g_o together constitute the decoder illustrated in Figure 1.

Our goal is to characterize the region of all achievable (R, D_s, D_o) tuples. Hence, we define the semantic rate distortion function as follows[2]:

$$R(D_s, D_o) = \inf\{R : (R, D_s, D_o) \text{ is achievable}\}. \tag{5}$$

Clearly, characterizing the semantic rate distortion function $R(D_s, D_o)$ is equivalent to characterizing the achievable region of (R, D_s, D_o).

We will also consider a variant of the distortion constraint; that is, the state distortion and the observation distortion are linearly combined to yield a single overall distortion. Hence, instead of (3) and (4), the decoding functions are required to satisfy the following weighted distortion constraint:

$$\mathbb{E}\left[w_s d_s(S^n, \hat{S}^n) + w_o d_o(X^n, \hat{X}^n)\right] \le \bar{D} + \epsilon, \tag{6}$$

where w_s and w_o are non-negative weighting coefficients.

It is also natural to generalize the system model to include several intrinsic state variables each associated with a specified reproduction and a distortion. Such a semantic source is described by a tuple of random variables, $(S_0, S_1, \ldots, S_{k-1}, X)$, with joint probability distribution $p(s_0, s_1, \ldots, s_{k-1}, x)$ over $\mathcal{S}_0 \times \mathcal{S}_1 \times \ldots \times \mathcal{S}_{k-1} \times \mathcal{X}$, where each S_j is an intrinsic state reflecting a certain semantic aspect of the source. The decoder now consists of an observation decoding function and k state decoding functions, among which $g_{s,j}$ maps the codeword $W \in \{1, 2, \ldots, 2^{\lfloor n(R+\epsilon)\rfloor}\}$ into a reproduction sequence $\hat{S}_j^n$ to satisfy

$$\mathbb{E}\left[d_{s,j}(S_j^n, \hat{S}_j^n)\right] \le D_{s,j} + \epsilon. \tag{7}$$

[2]This is the operational definition of a rate distortion function, which has been widely used (see, for example, [3], [28] [29]).

The notion of achievability can be defined in a similar fashion with respect to the tuple $(R, D_{s,0}, D_{s,1}, \ldots, D_{s,k-1}, D_o)$, and the semantic rate distortion function is consequently defined as

$$R(D_{s,0}, D_{s,1}, \cdots, D_{s,k-1}, D_o) = \inf\{R : (R, D_{s,0}, D_{s,1}, \cdots, D_{s,k-1}, D_o) \text{ is achievable}\}. \tag{8}$$

Examples of such semantic sources with multiple semantic aspects can be found in [5], [9], which consider a hierarchy of image or video features, each feature associated with a quality metric.

III. Semantic Rate Distortion Function

In this section, we establish in the following theorem a single-letter characterization of the semantic rate distortion function $R(D_s, D_o)$ defined in Section II.

Theorem 1: For a given semantic source (S, X) with $p(s, x)$ over $\mathcal{S}\times\mathcal{X}$, reproduction alphabet $\hat{\mathcal{S}}\times\hat{\mathcal{X}}$, and distortion measures d_s and d_o, the semantic rate distortion function $R(D_s, D_o)$ is as follows:

$$R(D_s, D_o) = \min_{p(\hat{s},\hat{x}|x)} I(X; \hat{S}, \hat{X}) \tag{9}$$

$$\text{s.t. } \mathbb{E}\left[d_o(X, \hat{X})\right] \le D_o, \tag{10}$$

$$\mathbb{E}\left[\hat{d}_s(X, \hat{S})\right] \le D_s, \tag{11}$$

where

$$\hat{d}_s(x, \hat{s}) = \mathbb{E}\left[d_s(S, \hat{s})|x\right] = \sum_{s\in\mathcal{S}} p(s|x) d_s(s, \hat{s}), \tag{12}$$

and $S, X, \hat{S}, \hat{X}$ constitute a Markov chain $S \leftrightarrow X \leftrightarrow (\hat{S}, \hat{X})$.

Proof: See Appendix I. □

Here we briefly discuss the basic idea of the proof of Theorem 1. There are two main ingredients in the problem formulation: an indirect rate distortion problem which has been studied in [30], [31] [32, Chap. 3, Sec. 5] [33], and a rate distortion problem with several distortion constraints which has been studied in [34, Sec. VII] [3, Prob. 10.19] [35, Prob. 7.14]. A key is to recognize reproducing $\hat{S}$ as an indirect rate distortion problem, for which the state distortion between S and $\hat{S}$ can be equivalently converted to a distortion between X and $\hat{S}$. Indeed, the converted distortion is nothing but the conditional expectation of the original state distortion $d_s(S, \hat{s})$, over $p(s|x)$. This conversion hence circumvents the difficulty due to the absence of access to S at the encoder. The detailed derivation, which is based on a unified treatment in [33], is given in Appendix I.

We note that the semantic rate distortion function can be non-trivial even for the special case where S is a deterministic function of X, because from a lossy reproduction of X it is generally impossible to reproduce S in a lossless fashion. Specifically, suppose that $S = g(X)$. Then $\hat{d}_s(x, \hat{s})$ can be simplified into

$$\hat{d}_s(x, \hat{s}) = \sum_{s\in\mathcal{S}} p(s|x) d_s(s, \hat{s}) = d_s(g(x), \hat{s}). \tag{13}$$

Similar to standard rate distortion functions, a corollary of the semantic rate distortion function as given by Theorem 1 is the following regarding monotonicity and convexity.

Corollary 1: The semantic rate distortion function $R(D_s, D_o)$ in Theorem 1 has the following properties:

- $R(D_s, D_o)$ is monotonically nonincreasing with D_s and D_o.
- $R(D_s, D_o)$ is jointly convex with respect to (D_s, D_o).
- The contour set $\{(D_s, D_o) : R(D_s, D_o) \le R\}$ is convex for any $R \ge 0$.

Proof: The proof of the first two properties is exactly the same as that for standard rate distortion functions; see, e.g., [3]. The third property is then an immediate corollary of the second property. □

Corollary 1 implies a trade-off between the two distortions: for a given code rate, the smaller the state distortion, the larger the observation distortion, and vice versa. Concrete numerical examples can be found in Section IV, where Figures 2 and 4 plot the achievable regions of (R, D_s, D_o) and their projections under different values of R, for two experimental setups, respectively. These plots demonstrate that for fixed R, the achievable (D_s, D_o) pairs form a convex region, whose boundary exhibits a trade-off between D_s and D_o. Hence a sensible coding scheme of a semantic source should exhibit such behavior.

Now consider the weighted distortion constraint (6). We have the following corollary.

Corollary 2: For a given semantic source under the weighted distortion constraint (6), the rate distortion function is as follows:

$$R(\bar{D}) = \min\left\{R(D_s, D_o) | w_s D_s + w_o D_o \le \bar{D}\right\}. \tag{14}$$

Proof: Given the semantic rate distortion function $R(D_s, D_o)$ in Theorem 1, we have that any coding scheme that achieves $(R, \bar{D})$ should achieve a (R, D_s, D_o) tuple for the semantic rate distortion problem under distortion constraints (3) and (4), for some D_s and D_o satisfying $w_s D_s + w_o D_o \le \bar{D}$, and vice versa. □

We end this section with the semantic rate distortion function (8) for semantic sources with several intrinsic states, as given by the following corollary. Its proof is essentially identical to that of Theorem 1.

Corollary 3: For a semantic source $(S_0, S_1, \ldots, S_{k-1}, X)$ with $p(s_0, s_1, \ldots, s_{k-1}, x)$ over $\mathcal{S}_0 \times \mathcal{S}_1 \times \ldots \times \mathcal{S}_{k-1} \times \mathcal{X}$, reproduction alphabet $\hat{\mathcal{S}}_0 \times \hat{\mathcal{S}}_1 \times \ldots \times \hat{\mathcal{S}}_{k-1} \times \hat{\mathcal{X}}$, and distortion measures $\{d_{s_j}\}_{j=0,1,\ldots,k-1}$ and d_o, the semantic rate distortion function $R(D_{s_0}, D_{s_1}, \ldots, D_{s_{k-1}}, D_o)$ is as follows:

$$R(D_{s_0}, D_{s_1}, \ldots, D_{s_{k-1}}, D_o) = \min_{p(\hat{s}_0,\hat{s}_1,\ldots,\hat{s}_{k-1},\hat{x}|x)} I(X; \hat{S}_0, \hat{S}_1, \ldots, \hat{S}_{k-1}, \hat{X}) \tag{15}$$

$$\text{s.t. } \mathbb{E}\left[d_o(X, \hat{X})\right] \le D_o, \tag{16}$$

$$\mathbb{E}\left[\hat{d}_{s_j}(X, \hat{S}_j)\right] \le D_{s_j}, \quad j = 0, 1, \ldots, k-1, \tag{17}$$

where

$$\hat{d}_{s_j}(x,\hat{s}_j)=\mathbb{E}\left[d_{s_j}(S_j,\hat{s}_j)|x\right]=\sum_{s_j\in\mathcal{S}_j}p(s_j|x)d_{s_j}(s_j,\hat{s}_j), \tag{18}$$

and $\{S_j\}_{j=0,1,\ldots,k-1},X,\{\hat{S}_j\}_{j=0,1,\ldots,k-1},\hat{X}$ constitute a Markov chain $(S_0,S_1,\ldots,S_{k-1})\leftrightarrow X\leftrightarrow(\hat{S}_0,\hat{S}_1,\ldots,\hat{S}_{k-1},\hat{X})$.

IV. Gaussian Observation With Linear State-Observation Relationship

Theorem 1 establishes the general form of the semantic rate distortion function, which comes with an optimization problem, extending its counterpart in a standard rate distortion problem. In this section, we specialize the general result to a case where the extrinsic observation X is Gaussian and the intrinsic state-extrinsic observation pair (S,X) satisfies a linear relationship, under quadratic distortion measures.

The extrinsic observation X obeys a multivariate Gaussian distribution $\mathcal{N}(0,\mathbf{K}_X)$,[3] where $\mathbf{K}_X$ is an $m\times m$ positive semi-definite matrix. The intrinsic state S is given by

$$S=\mathbf{H}X+Z, \tag{19}$$

where $\mathbf{H}$ is an $l\times m$ matrix, and Z is a random vector independent of X, with zero mean and covariance matrix $\mathbf{K}_Z$. Note that we neither restrict Z to be Gaussian nor require $\mathbf{H}$ or $\mathbf{K}_Z$ to be full-rank. According to (19), the intrinsic state S is a linear transformation of X, further disturbed by an independent component Z. This linear assumption holds for jointly Gaussian intrinsic state S and extrinsic observation X, and can usually be extended to non-Gaussian models as well, either precisely or approximately, for example, when a linear estimator of S conditioned upon X can be obtained by traditional statistical methods, or by multilayer perceptron (MLP) neural networks alternatively [36]. On the other hand, note that the linear assumption no longer holds when one invokes nonlinear mappings, and deriving an analytical form of the corresponding semantic rate distortion function will generally be an extremely difficult task.

This model covers the special case where (S,X) are jointly Gaussian. In fact, if (S,X) are jointly Gaussian with zero mean and covariance matrix

$$\begin{bmatrix}\mathbf{K}_S & \mathbf{K}_{SX}\\ \mathbf{K}_{SX}^T & \mathbf{K}_X\end{bmatrix}, \tag{20}$$

we can represent S according to

$$S=\mathbf{K}_{SX}\mathbf{K}_X^{-1}X+Z, \tag{21}$$

where $Z\sim\mathcal{N}(0,\mathbf{K}_S-\mathbf{K}_{SX}\mathbf{K}_X^{-1}\mathbf{K}_{SX}^T)$; that is, $\mathbf{H}=\mathbf{K}_{SX}\mathbf{K}_X^{-1}$ and $\mathbf{K}_Z=\mathbf{K}_S-\mathbf{K}_{SX}\mathbf{K}_X^{-1}\mathbf{K}_{SX}^T$.

We consider quadratic distortion measures, defined as

$$d_s(s,\hat{s})=\|s-\hat{s}\|_2^2=\mathrm{tr}(s-\hat{s})(s-\hat{s})^T, \tag{22}$$

$$d_o(x,\hat{x})=\|x-\hat{x}\|_2^2=\mathrm{tr}(x-\hat{x})(x-\hat{x})^T. \tag{23}$$

Consequently, we have

$$\mathbb{E}\left[d_s(S,\hat{S})\right]=\mathrm{tr}(\mathbf{K}_{S-\hat{S}}), \tag{24}$$

$$\mathbb{E}\left[d_o(X,\hat{X})\right]=\mathrm{tr}(\mathbf{K}_{X-\hat{X}}). \tag{25}$$

[3]We use $\mathbf{K}_V$ to denote the covariance matrix of a random column vector V.

For the considered model (19), we can derive its semantic rate distortion function, given by the following theorem.

Theorem 2: The semantic rate distortion function for the semantic source with Gaussian extrinsic observation and linear state-observation relationship (19), under quadratic distortion measures (22) and (23), is given by:

$$R_{\mathcal{G}}(D_s,D_o)=\min_{\mathbf{\Delta}\in\mathcal{S}_m}\ \frac{1}{2}\log\left(\frac{\det(\mathbf{K}_X)}{\det(\mathbf{\Delta})}\right) \tag{26}$$

$$\text{s.t. }\mathbf{O}\prec\mathbf{\Delta}\preceq\mathbf{K}_X, \tag{27}$$

$$\mathrm{tr}(\mathbf{H}\mathbf{\Delta}\mathbf{H}^T)\le D_s-\mathrm{tr}(\mathbf{K}_Z), \tag{28}$$

$$\mathrm{tr}(\mathbf{\Delta})\le D_o. \tag{29}$$

where $\mathcal{S}_m$ denotes the set of all $m\times m$ positive definite matrices. Note that here we use a subscript $\mathcal{G}$ to emphasize that the extrinsic observation is Gaussian.

Proof: See Appendix II. □

From (28), when Z is sufficiently strong so that $\mathrm{tr}(\mathbf{K}_Z)>D_s$, the optimization (26) is no longer feasible and hence $R_{\mathcal{G}}(D_s,D_o)=\infty$. Otherwise, there is no further restriction on $\mathbf{K}_Z$. For example, even if $Z=0$, i.e., the relationship between S and X is deterministic as $S=\mathbf{H}X$, the optimization problem in Theorem 2 is still non-trivial.

A simplified case arises when $\mathbf{H}$ is an orthogonal matrix satisfying $\mathbf{H}^T\mathbf{H}=\mathbf{I}$. In this case, (28) becomes

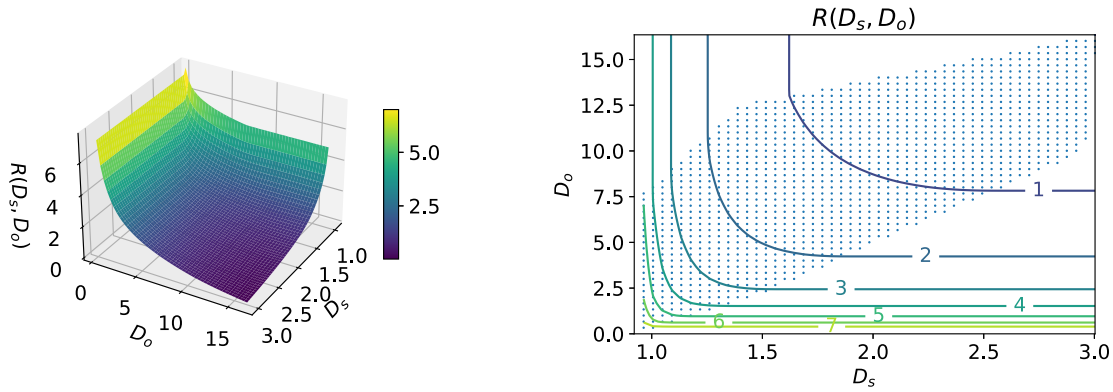
$$\mathrm{tr}(\mathbf{H}\mathbf{\Delta}\mathbf{H}^T)=\mathrm{tr}(\mathbf{\Delta}\mathbf{H}^T\mathbf{H})=\mathrm{tr}(\mathbf{\Delta})\le D_s-\mathrm{tr}(\mathbf{K}_Z), \tag{30}$$

which can then be combined with (29) leading to a single distortion constraint

$$\mathrm{tr}(\mathbf{\Delta})\le\min\{D_o,D_s-\mathrm{tr}(\mathbf{K}_Z)\}. \tag{31}$$

In Theorem 2, the matrix $\mathbf{\Delta}$ which we optimize corresponds to the mean squared error (MSE) of estimating X based upon $\hat{X}$ at the decoder. The key to the proof of Theorem 2 is to show that the semantic rate distortion function is achieved by a Gaussian reproduction. This is similar to situations in several Gaussian lossy compression problems, including the standard Gaussian rate distortion problem [12] and the Gaussian quadratic CEO problem [37]. Existing techniques based on the entropy power inequality (EPI), extremal inequalities, and Fisher information inequalities may also be interpreted as the optimality of Gaussian reproduction for the minimum mean squared error (MMSE) estimation under a given MSE constraint. In our analysis, we further need to accommodate with two MSE constraints, corresponding to the intrinsic state and the extrinsic observation, respectively.

Compared with the general form of semantic rate distortion function in Theorem 1, Theorem 2 involves only one matrix-valued optimization variable $\mathbf{\Delta}$, which, as remarked in the previous paragraph, is the MSE of estimating X based upon $\hat{X}$ alone. In fact, the solution exhibits a Markov structure, i.e., $S\leftrightarrow X\leftrightarrow\hat{X}\leftrightarrow\hat{S}$. To help understand the optimality of the Markov chain solution, supposing that an alternative solution $(\hat{X}',\hat{S}')$ is given which does not satisfy the Markov structure, consequently one can form an improved reproduction as $\hat{X}=\mathbb{E}(X|\hat{X}',\hat{S}')$, satisfying the Markov structure and achieving the same code rate $I(X;\hat{X},\hat{S}')=I(X;\hat{X}',\hat{S}')$.



(a) Surface plot of $R_{\mathcal{G}}(D_s, D_o)$ (b) Contour plot of $R_{\mathcal{G}}(D_s, D_o)$

Fig. 2. Surface and contour plots of the semantic rate distortion function $R_{\mathcal{G}}(D_s, D_o)$ for the toy example.

The Markov chain solution further suggests a "two-stage" coding interpretation which is in fact extensively adopted in practice: the decoder first generates a reproduction for X as $\hat{X}$, and then uses that reproduction to further generate a reproduction for S as $\hat{S}$. Similar to the standard Gaussian rate distortion problem, the optimal $\hat{X}$ can be constructed with the aid of a "test channel", for which $\hat{X}$ as the channel input is Gaussian and the additive Gaussian noise of the test channel has a covariance matrix Δ, thereby producing X as the desired channel output. To generate $\hat{S}$ based upon $\hat{X}$, it suffices to adopt a linear transform $\hat{S} = \mathbf{H}\hat{X}$. On the other hand, the Markov chain solution does not mean that the reproduction of S is trivial, because the fidelity criterion on X still needs to be adjusted according to D_s. The detailed arguments are given in the proof in Appendix II.

An interesting property of the semantic rate distortion function derived in Theorem 2 is that it is in fact an upper bound for all semantic sources with the same covariance structure under the quadratic distortion measure. This essentially indicates that a semantic source with Gaussian extrinsic observation is the hardest to describe, analogous to its counterpart in conventional source coding problems (see, e.g., [3, Exercise 10.8]). Formally, we have the following corollary.

Corollary 4: For a semantic source (S, X) with general probability density function, whose covariance matrix is given by (20), its semantic rate distortion function subject to quadratic distortion constraints (22) and (23) satisfies $R(D_s, D_o) \leq R_{\mathcal{G}}(D_s, D_o)$, where $R_{\mathcal{G}}(D_s, D_o)$ is the semantic rate distortion function given in Theorem 2, with $\mathbf{H} = \mathbf{K}_{SX}\mathbf{K}_X^{-1}$ and $\mathbf{K}_Z = \mathbf{K}_S - \mathbf{K}_{SX}\mathbf{K}_X^{-1}\mathbf{K}_{SX}^T$.

Proof: See Appendix III. □

A. Computation of the Semantic Rate Distortion Function

We remark that the optimization problem in Theorem 2 is convex, and hence can be numerically solved by software like CVX in an efficient and stable fashion. In this subsection we present some illustrative numerical examples.

Our first example is a small-scale toy model, given by

$$\mathbf{K}_X = \begin{bmatrix} 11 & 0 & 0.5 \\ 0 & 3 & -2 \\ 0.5 & -2 & 2.35 \end{bmatrix},$$

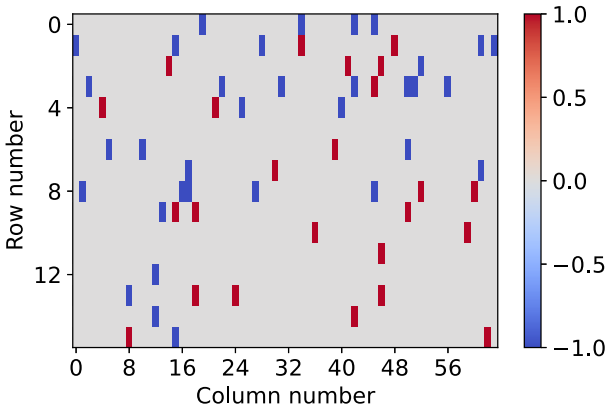


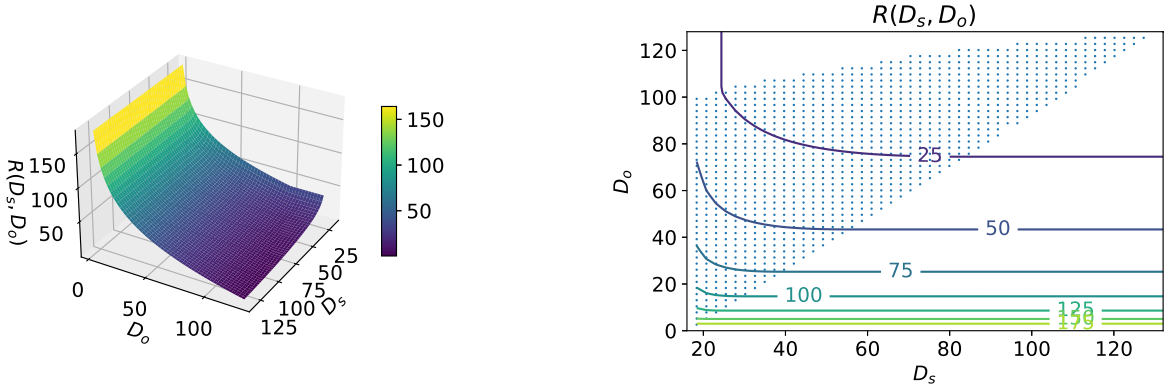
Fig. 3. A 16×64 transformation matrix $\mathbf{H}$ shown as a two-dimensional grid. Elements are shown as cells with different colors corresponding to their values: blue for -1, red for 1, and gray for 0.

$$\mathbf{H} = \begin{bmatrix} 0.0701 & 0.305 & 0.457 \\ -0.0305 & -0.220 & 0.671 \end{bmatrix},$$
$$\mathbf{K}_Z = \begin{bmatrix} 0.701 & -0.305 \\ -0.305 & 0.220 \end{bmatrix}.$$

The resulting semantic rate distortion function is computed as displayed in Figure 2. The dotted region in Figure 2(b) indicates that both constraints (28) and (29) are active. The trade-off between the two distortions are clear: the smaller the state distortion, the larger the observation distortion, and vice versa.

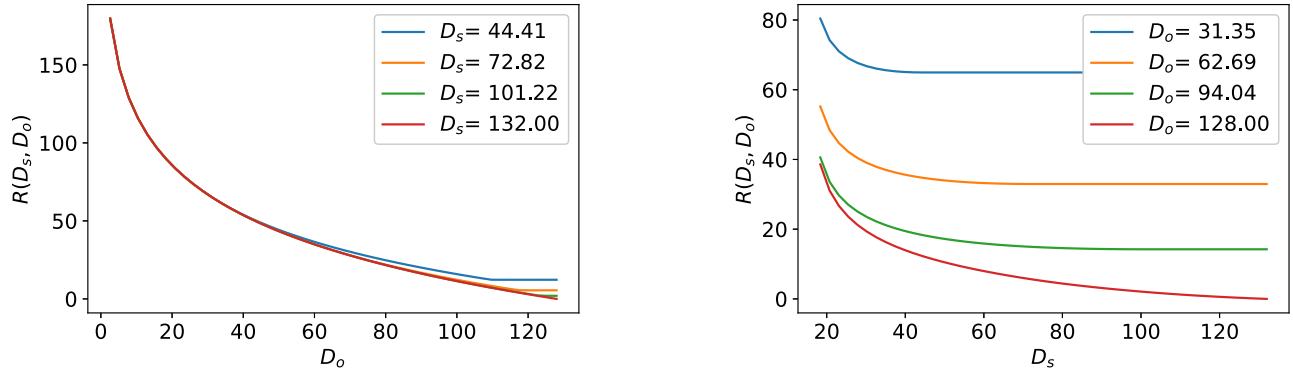
Our second example captures a sparse state-observation relationship, as follows. The extrinsic observation is a length-64 vector $X = [X_1, \cdots, X_{64}]^T$ consisting of i.i.d. $\mathcal{N}(0, 2)$ random variables. The transformation matrix $\mathbf{H}$ is a randomly masked 16×64 Rademacher matrix; that is, we first generate a Rademacher matrix whose elements are i.i.d. taking values $\{1, -1\}$ with equal probability $1/2$, and then independently reset these elements to zero with probability 0.95. A realization of $\mathbf{H}$ is shown in Figure 3. The noise vector $Z = [Z_1, \cdots, Z_{16}]^T$ consists of i.i.d. $\mathcal{N}(0, 1)$ random variables.

We numerically solve the semantic rate distortion function according to Theorem 2, and a typical surface of $R_{\mathcal{G}}(D_s, D_o)$ is illustrated in Figure 4(a). More details of $R_{\mathcal{G}}(D_s, D_o)$ can be seen from the contour plot in Figure 4(b), wherein the dotted region indicate that both constraints (28) and (29)



(a) Surface plot of $R_{\mathcal{G}}(D_s, D_o)$ (b) Contour plot of $R_{\mathcal{G}}(D_s, D_o)$

Fig. 4. Surface and contour plots of the semantic rate distortion function $R_{\mathcal{G}}(D_s, D_o)$ for the example of a sparse state-observation relationship.



(a) $R_{\mathcal{G}}(D_s, D_o)$ versus D_o (b) $R_{\mathcal{G}}(D_s, D_o)$ versus D_s

Fig. 5. The semantic rate distortion function $R_{\mathcal{G}}(D_s, D_o)$ as a function of D_o or D_s.

are active. From Figure 4(b), it is evident that describing the extrinsic observation X tends to be much more costly than describing the intrinsic state S: at the same code rate, the achieved D_s is generally much lower than the achieved D_o.

Another interesting fact regarding $R_{\mathcal{G}}(D_s, D_o)$ can be inferred from the dotted region in the contour plot Figure 4(b), and is more clearly revealed by plotting the trends of $R_{\mathcal{G}}(D_s, D_o)$ as a function of D_o (for fixed D_s) or D_s (for fixed D_o), shown in Figures 5(a) and 5(b), respectively. We find that, the code rate $R_{\mathcal{G}}(D_s, D_o)$ as a function of D_o does not seem to be sensitive to the choice of D_s. This fact has an important consequence for designing lossy compression schemes for semantic sources: although several different codes may have similar performance in terms of reproducing the extrinsic observation, they can differ considerably in terms of reproducing the intrinsic state. A heuristic explanation is as follows: since X is a high-dimensional vector, describing it along several different directions may lead to similar quadratic distortion performance; but since S corresponds to a low-dimensional feature of X, its reproduction only favors the direction of describing X that retains the feature of S the best.

B. Generalizations of Theorem 2

We can derive from Theorem 2 several corollaries corresponding to the variants of the problem formulation in Section II.

First, let us consider replacing the quadratic distortion measures by the positive semi-definite distortion constraints. Following the same arguments in the proof of Theorem 2, we again arrive at the optimality of Gaussian descriptions under positive semi-definite distortion constraints, and hence the following corollary characterizes the semantic rate distortion function.

Corollary 5: Consider the positive semi-definite distortion measures as

$$d_s(s, \hat{s}) = (s - \hat{s})(s - \hat{s})^T,$$
$$d_o(x, \hat{x}) = (x - \hat{x})(x - \hat{x})^T.$$

The semantic rate distortion function is given by

$$R(\mathbf{D}_s, \mathbf{D}_o) = \min_{\mathbf{\Delta} \in \mathcal{S}_m} \frac{1}{2} \log \left(\frac{\det(\mathbf{K}_X)}{\det(\mathbf{\Delta})} \right) \tag{32}$$
$$\text{s.t. } \mathbf{O} \prec \mathbf{\Delta} \preceq \mathbf{K}_X, \tag{33}$$
$$\mathbf{H}\mathbf{\Delta}\mathbf{H}^T \preceq \mathbf{D}_s - \mathbf{K}_Z, \tag{34}$$
$$\mathbf{\Delta} \preceq \mathbf{D}_o. \tag{35}$$

This is a semi-definite programming problem and can be readily solved by software.

Now consider the weighted distortion constraint, where the distortion measure is defined as a weighted sum of two individual distortion measures, i.e.

$$\bar{d} = w_s d_s(s, \hat{s}) + w_o d_o(x, \hat{x})$$
$$= w_s \|s - \hat{s}\|_2^2 + w_o \|x - \hat{x}\|_2^2. \tag{36}$$

Applying Corollary 2, we obtain the semantic rate distortion function in the following corollary.

Corollary 6: For the weighted distortion measure $\bar{d}$, the semantic rate distortion function $R(\bar{D})$ is given by

$$R(\bar{D}) = \min_{\mathbf{\Delta}\in\mathcal{S}_m} \frac{1}{2}\log\left(\frac{\det(\mathbf{K}_X)}{\det(\mathbf{\Delta})}\right) \tag{37}$$

$$\text{s.t. } \mathbf{O} \prec \mathbf{\Delta} \preceq \mathbf{K}_X, \tag{38}$$

$$\operatorname{tr}((w_s\mathbf{H}^T\mathbf{H} + w_o\mathbf{I}_m)\mathbf{\Delta}) \le \bar{D} - w_s\operatorname{tr}(\mathbf{K}_Z). \tag{39}$$

Finally, consider the case of k intrinsic states. The extrinsic observation X is still $\mathcal{N}(0, \mathbf{K}_X)$. For each $j \in \{0, 1, \cdots, k-1\}$, the j-th intrinsic state is generated according to

$$S_j = \mathbf{H}_j X + Z_j,$$

where $\mathbf{H}_j$ is an $l_j \times m$ matrix, and Z_j is a random vector independent of X, with zero mean and covariance matrix $\mathbf{K}_{Z_j}$. We consider quadratic distortion measures, as

$$d_{s_j}(s_j, \hat{s}_j) = \|s_j - \hat{s}_j\|_2^2, \quad j = 0, 1, \ldots, k-1, \tag{40}$$

$$d_o(x, \hat{x}) = \|x - \hat{x}\|_2^2. \tag{41}$$

The semantic rate distortion function is given by the following corollary.

Corollary 7: For the semantic source with a Gaussian extrinsic observation and k intrinsic states, the semantic rate distortion function under distortion measures d_{s_0}, d_{s_1}, $\cdots$, $d_{s_{k-1}}$, d_o is

$$\begin{aligned} &R(D_{s_0}, D_{s_1}, \cdots, D_{s_{k-1}}, D_o) \\ &\quad= \min_{\mathbf{\Delta}\in\mathcal{S}_m} \frac{1}{2}\log\left(\frac{\det(\mathbf{K}_X)}{\det(\mathbf{\Delta})}\right) \\ &\qquad\text{s.t. } \mathbf{O} \prec \mathbf{\Delta} \preceq \mathbf{K}_X, \\ &\qquad\quad \operatorname{tr}(\mathbf{H}_j\mathbf{\Delta}\mathbf{H}_j^T) \le D_{s_j} - \operatorname{tr}(\mathbf{K}_{Z_j}), \\ &\qquad\qquad\qquad j \in \{0, 1, \cdots, k-1\}, \\ &\qquad\quad \operatorname{tr}(\mathbf{\Delta}) \le D_o. \end{aligned}$$

V. Weighted Reverse Water-Filling

Analogous to the standard Gaussian rate distortion problem wherein (after appropriate linear transformation) the solution can be interpreted as a reverse water-filling type of rate allocation, for the semantic rate distortion function in Theorem 2, under a diagonalizability condition, the solution can also be interpreted as reverse water-filling, but with appropriately weighted water levels.

For the model of Gaussian observation with linear state-observation relationship in Section IV, we further assume that the following diagonalizability condition is satisfied: there exists an unitary matrix $\mathbf{Q}$ such that

- $\mathbf{Q}^\dagger\mathbf{K}_X\mathbf{Q} = \operatorname{diag}(\sigma_1, \sigma_2, \cdots, \sigma_m)$,
- $\mathbf{Q}^\dagger\mathbf{H}^T\mathbf{H}\mathbf{Q} = \operatorname{diag}(\alpha_1, \alpha_2, \cdots, \alpha_m)$

simultaneously hold. Here it loses no generality to order $\{\alpha_i\}_{i=1}^m$ so that $\alpha_1 \ge \alpha_2 \ge \cdots \ge \alpha_m$. Denoting the rank of $\mathbf{H}^T\mathbf{H}$ as $q \le m$, then $\alpha_q > 0$ and $\alpha_{q+1} = \cdots = \alpha_m = 0$.

Lemma 1: Under the diagonalizability condition, the resulting optimal $\mathbf{\Delta}$ takes the form

$$\mathbf{\Delta} = \mathbf{Q}\operatorname{diag}(\delta_1, \delta_2, \cdots, \delta_m)\mathbf{Q}^\dagger, \tag{42}$$

and the semantic rate distortion function in Theorem 2 can be further written in terms of the following optimization problem:

$$R_{\mathcal{G}}(D_s, D_o) = \min_{\delta_1,\delta_2,\cdots,\delta_m} \frac{1}{2}\sum_{j=1}^m \log\left(\frac{\sigma_j}{\delta_j}\right) \tag{43}$$

$$\text{s.t. } 0 < \delta_j \le \sigma_j, \quad \forall j \in \{1, 2, \cdots, m\}, \tag{44}$$

$$\sum_{j=1}^m \alpha_j\delta_j \le D_s - \operatorname{tr}(\mathbf{K}_Z), \tag{45}$$

$$\sum_{j=1}^m \delta_j \le D_o. \tag{46}$$

Proof: See Appendix IV. □

In order to describe the weighted reverse water-filling solution, we first introduce the following curves.

- Curve C_s:

$$C_s = \left\{ \begin{bmatrix} \sum_{j=1}^m \alpha_j \min\left(\sigma_j, \frac{1}{\lambda}\right) + \operatorname{tr}(\mathbf{K}_Z) \\ \sum_{j=1}^m \min\left(\sigma_j, \frac{1}{\lambda}\right) \end{bmatrix} \middle| \lambda > 0 \right\}, \tag{47}$$

  which starts from $(\operatorname{tr}(\mathbf{H}\mathbf{K}_X\mathbf{H}^T + \mathbf{K}_Z), \operatorname{tr}(\mathbf{K}_X))$ and ends at $(\operatorname{tr}(\mathbf{K}_Z), 0)$.
- Curve C_o:

$$C_o = \left\{ \begin{bmatrix} \sum_{j=1}^q \alpha_j \min\left(\sigma_j, \frac{1}{\mu\alpha_j}\right) + \operatorname{tr}(\mathbf{K}_Z) \\ \sum_{j=1}^q \min\left(\sigma_j, \frac{1}{\mu\alpha_j}\right) + \sum_{j=q+1}^m \sigma_j \end{bmatrix} \middle| \mu > 0 \right\}, \tag{48}$$

  which starts from $(\operatorname{tr}(\mathbf{H}\mathbf{K}_X\mathbf{H}^T + \mathbf{K}_Z), \operatorname{tr}(\mathbf{K}_X))$ and ends at $(\operatorname{tr}(\mathbf{K}_Z), \sum_{j=q+1}^m \sigma_j)$. Here, $\sum_{j=1+1}^m \sigma_j$ is interpreted as 0 if $\mathbf{H}^T\mathbf{H}$ is full-rank and thus $q = m$.

We then introduce the following partitioning of the (D_s, D_o) plane, based upon the curves C_s and C_o:

- $A_0 = \{(D_s, D_o) | D_s \ge \operatorname{tr}(\mathbf{H}\mathbf{K}_X\mathbf{H}^T + \mathbf{K}_Z), D_o \ge \operatorname{tr}(\mathbf{K}_X)\}$;
- A_1: on the right of the curve C_s, and between the two horizontal lines $D_o = 0$ and $D_o = \operatorname{tr}(\mathbf{K}_X)$;
- A_2: above the curve C_o, and between the two vertical lines $D_s = \operatorname{tr}(\mathbf{K}_Z)$ and $D_s = \operatorname{tr}(\mathbf{H}\mathbf{K}_X\mathbf{H}^T + \mathbf{K}_Z)$;
- A_3: surrounded by the curves C_s and C_o, and the vertical line $D_s = \operatorname{tr}(\mathbf{K}_Z)$.

An example of the partitioning above is plotted in Figure 6.

The following theorem describes the weighted reverse water-filling solution.

Theorem 3: For the model of Gaussian observation with linear state-observation relationship in Section IV, under the diagonalizability condition, the optimal $\mathbf{\Delta} = \mathbf{Q}\operatorname{diag}(\delta_1, \delta_2, \cdots \delta_m)\mathbf{Q}^\dagger$ is given by

TABLE I
ACTIVITY OF CONSTRAINTS (45) AND (46) IN A_0, A_1, A_2 AND A_3

	(45) active	(45) inactive
(46) inactive	A_2	A_0
(46) active	A_3	A_1

- If $(D_s, D_o) \in A_0$:

$$\delta_j^* = \sigma_j, \quad \forall j \in \{1, 2, \cdots, m\}. \tag{49}$$

- If $(D_s, D_o) \in A_1$:

$$\delta_j^* = \min\left(\sigma_j, \frac{1}{\lambda}\right), \quad \forall j \in \{1, 2, \cdots, m\}, \tag{50}$$

where λ is chosen to satisfy $\sum_{j=1}^m \delta_j^* = D_o$.

- If $(D_s, D_o) \in A_2$:

$$\delta_j^* = \begin{cases} \min\left(\sigma_j, \dfrac{1}{\mu\alpha_j}\right), & \alpha_j > 0 \\ \sigma_j, & \alpha_j = 0, \end{cases} \quad \forall j \in \{1, 2, \cdots, q\}, \tag{51}$$

where μ is chosen to satisfy $\sum_{j=1}^q \alpha_j \delta_j^* = D_s - \mathrm{tr}(\mathbf{K}_Z)$.

- If $(D_s, D_o) \in A_3$:

$$\delta_j^* = \min\left(\sigma_j, \frac{1}{\lambda + \mu\alpha_j}\right), \quad \forall j \in \{1, 2, \cdots, m\}, \tag{52}$$

where λ, μ are chosen to satisfy $\sum_{j=1}^m \delta_j^* = D_o$ and $\sum_{j=1}^q \alpha_j \delta_j^* = D_s - \mathrm{tr}(\mathbf{K}_Z)$.

Proof: See Appendix IV. □

The partitioning $\{A_0, A_1, A_2, A_3\}$ is closely related to activity of the constraints (45) and (46), as summarized in Table I. In A_0, both constraints are inactive, and hence the optimization is unconstrained yielding the trivial solution (49). In A_1, only the observation distortion constraint is active, and the solution (50) is a standard reverse water-filling with water level $1/\lambda$. In A_2, only the state distortion is active, and the solution (51) essentially makes the weighted eigenvalues $\alpha_1\delta_1$, $\alpha_2\delta_2$, $\cdots$, $\alpha_m\delta_m$ fulfill a reverse water-filling structure, with water level $1/\mu$. Alternatively, we may view the term $1/(\mu\alpha_j)$ in (51) as a water level with weight $1/\alpha_j$. In A_3, both constraints are active, and the solution (52) also fulfills a reverse water-filling structure with unequal water levels.

A. Case Study: Circulant $\mathbf{K}_X$ and $\mathbf{H}$ and Weighted Reverse Water-Filling in Frequency Domain

A case of special interest is where $\mathbf{K}_X$ and $\mathbf{H}$ are both circulant matrices [38]. As the dimension of X grows large, this models the scenario where X is a circularly stationary Gaussian process,[4] and S is obtained via passing X through a time-invariant linear filter whose response is given by the first row of $\mathbf{H}$. For a circulant matrix, the corresponding unitary matrix $\mathbf{Q}$ is the well known discrete Fourier transform (DFT) matrix, and its eigenvalues are the DFT of the first row of the matrix. Hence the weighted reverse water-filling may be interpreted as exercised in the frequency domain, similar to its counterpart for the standard rate distortion function of stationary Gaussian processes.

[4]If we remove the circulant restriction and consider a stationary Gaussian process, then we encounter a Toeplitz $\mathbf{K}_X$, for which our solution still approximately applies; see, e.g., [38].

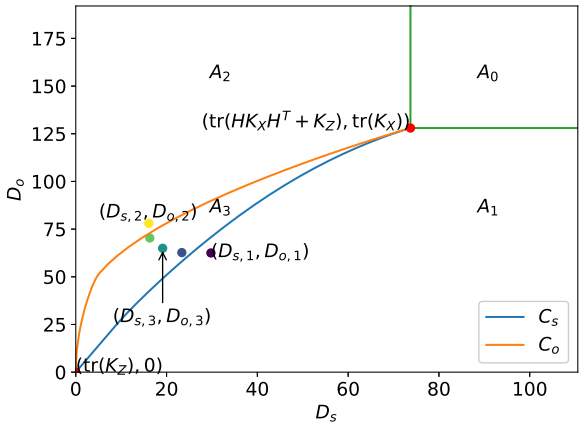


Fig. 6. The (D_s, D_o) plane is divided into four regions A_0, A_1, A_2, A_3, which determine the form of the optimal $\mathbf{\Delta}$. Five points on the contour $R_{\mathcal{G}}(D_s, D_o) = 50$ are marked with colors varying from purple to yellow.

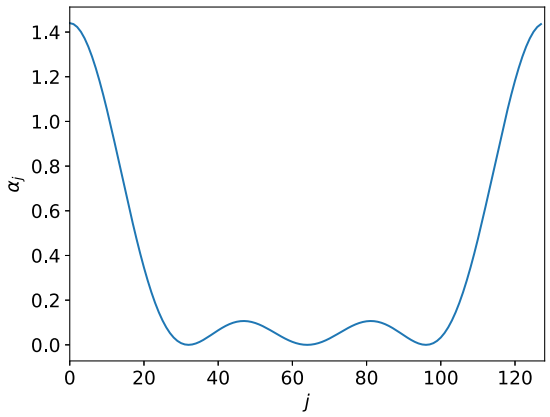


Fig. 7. Diagonal elements α_0, α_1, $\cdots$, α_{127} of $\mathbf{Q}^\dagger\mathbf{H}^T\mathbf{H}\mathbf{Q}$.

In the illustrative example below, consider $\mathbf{K}_X$ as a 128×128 circulant matrix with the first row

$$[1, 0.4, 0, \cdots, 0, 0.4],$$

$\mathbf{H}$ as a 128×128 circulant matrix with the first row

$$[0.3, 0.3, 0.3, 0.3, 0, \cdots, 0],$$

and $\mathbf{K}_Z$ as a 128×128 zero matrix (i.e., no noise in the state-observation relationship). Therefore, $\mathbf{Q}$ is the 128×128 DFT matrix whose (i, j)-th element is

$$\frac{1}{\sqrt{128}} e^{-\mathrm{i}\frac{2\pi}{128}ij}, \quad i, j = 0, 1, \cdots, 127.$$

The diagonal elements α_0, α_1, $\cdots$, α_{127} of $\mathbf{Q}^\dagger\mathbf{H}^T\mathbf{H}\mathbf{Q}$ are shown in Figure 7, and the diagonal elements σ_0, σ_1, $\cdots$, σ_{127} of $\mathbf{Q}^\dagger\mathbf{K}_X\mathbf{Q}$ are shown as the blue solid curve in Figure 8. Figure 6 shows the four regions A_0, A_1, A_2, A_3 and the two curves C_s, C_o. It also displays five points on the contour of $R_{\mathcal{G}}(D_s, D_o) = 50$, marked with colors varying from purple to yellow. The weighted reverse water-filling solution

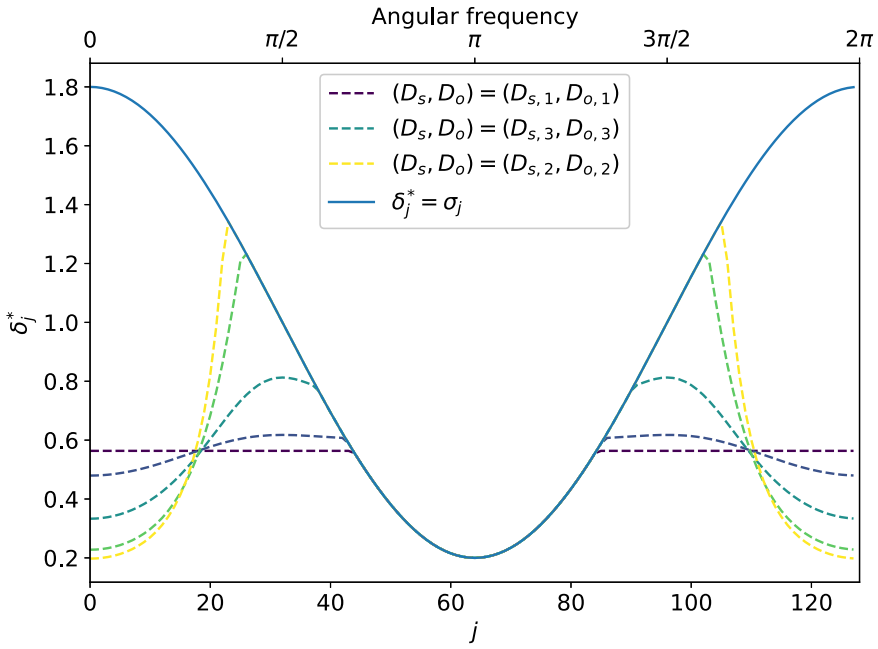


Fig. 8. Optimal diagonal elements $(\delta_1^*, \delta_2^*, \delta_3^*)$ of $\mathbf{Q}\mathbf{\Delta}\mathbf{Q}^T$ for the marked points in Figure 6, plotted with the colors in Figure 6.

$(\delta_0^*, \delta_1^*, \cdots, \delta_{127}^*)$ for these points are depicted in Figure 8. For $(D_{s,1}, D_{o,1})$, the optimal solution degenerates into a standard reverse water-filling form, as indicated by the purple line. When we go from $(D_{s,1}, D_{o,1})$ to $(D_{s,2}, D_{o,2})$, the water level begins to "ripple". Note that this weighted reverse water-filling can be viewed as exercised in the frequency domain, and the angular frequencies are marked on the top of Figure 8.

VI. Conclusion

We have provided a general source model to describe information sources that have semantic aspects, and proposed a corresponding rate distortion problem formulation for characterizing the amount of information content of such semantic sources. We have studied the case of Gaussian extrinsic observation subject to a linear state-observation relationship and a quadratic distortion structure. There are a variety of issues that we have not touched upon in the present work. First, calculating and bounding the semantic rate distortion functions for other interesting cases would make further use of our proposed framework, for example, when the intrinsic state is a discrete categorical random variable, corresponding to the important problem of classification; see [1] for some preliminary results. Second, a more challenging problem is to estimate the semantic rate distortion function, and more importantly, to develop effective lossy compression methods when the joint probability distribution of the intrinsic state and the extrinsic observation is not perfectly known, say, when only finite training data of the state-observation pair are available.

Appendix I
Proof of Theorem 1

The key to proving Theorem 1 is converting the semantic rate distortion problem into an equivalent standard rate distortion problem, with an indirect (state) distortion constraint and a direct (observation) distortion constraint. More precisely, we need to show that the constraint with respect to the state distortion measure $d_s(s, \hat{s})$ is equivalent to a constraint on a converted distortion measure $\hat{d}_s(x, \hat{s})$; that is, as long as a reproduction $\hat{S}$ satisfies the constraint on $\hat{d}_s(x, \hat{s})$, it will satisfy the constraint on $d_s(s, \hat{s})$, and vice versa.

A general and unified approach to the indirect rate-distortion function put forward in [33] is first showing that the one-shot expected distortion $\mathbb{E}\left[d_s(S, \hat{S})\right]$ is equivalent to $\mathbb{E}\left[\hat{d}_s(X, \hat{S})\right]$, and then invoking a tensorization argument to extend the one-shot equivalence to block codes. Here we directly illustrate how this can be accomplished for $S^n \leftrightarrow X^n \leftrightarrow (\hat{S}^n, \hat{X}^n)$ generated by an arbitrary encoder-decoder pair, as follows:

$$
\begin{aligned}
&\mathbb{E}\left[d_s(S^n, \hat{S}^n)\right] \\
&= \sum_{s^n, \hat{s}^n} p(s^n, \hat{s}^n) d_s(s^n, \hat{s}^n) \\
&= \sum_{s^n, x^n, \hat{s}^n} p(s^n, x^n, \hat{s}^n) d_s(s^n, \hat{s}^n) \\
&\overset{(a)}{=} \sum_{s^n, x^n, \hat{s}^n} p(s^n, x^n) p(\hat{s}^n | x^n) d_s(s^n, \hat{s}^n) \\
&= \sum_{x^n, \hat{s}^n} p(\hat{s}^n | x^n) \sum_{s^n} p(s^n, x^n) d_s(s^n, \hat{s}^n) \\
&\overset{(b)}{=} \sum_{x^n, \hat{s}^n} p(\hat{s}^n | x^n) \sum_{s^n} p(s^n, x^n) \frac{1}{n} \sum_{i=1}^{n} d_s(s_i, \hat{s}_i) \\
&= \sum_{x^n, \hat{s}^n} p(\hat{s}^n | x^n) \frac{1}{n} \sum_{i=1}^{n} \sum_{s^n} p(s^n, x^n) d_s(s_i, \hat{s}_i)
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(c)}{=} \sum_{x^n,\hat{s}^n} p(\hat{s}^n|x^n)\frac{1}{n}\sum_{i=1}^{n}\sum_{\bar{s}_i}\sum_{s_i\in\mathcal{S}} p(\bar{s}_i,\bar{x}_i)p(s_i,x_i)d_s(s_i,\hat{s}_i)\\
&= \sum_{x^n,\hat{s}^n} p(\hat{s}^n|x^n)\frac{1}{n}\sum_{i=1}^{n}\sum_{\bar{s}_i} p(\bar{s}_i,\bar{x}_i)\sum_{s_i\in\mathcal{S}} p(s_i,x_i)d_s(s_i,\hat{s}_i)\\
&= \sum_{x^n,\hat{s}^n} p(\hat{s}^n|x^n)\frac{1}{n}\sum_{i=1}^{n} p(\bar{x}_i)\sum_{s_i\in\mathcal{S}} p(s_i,x_i)d_s(s_i,\hat{s}_i)\\
&= \sum_{x^n,\hat{s}^n} p(\hat{s}^n|x^n)\frac{1}{n}\sum_{i=1}^{n} p(x^n)\sum_{s_i\in\mathcal{S}} p(s_i|x_i)d_s(s_i,\hat{s}_i)\\
&\overset{(d)}{=} \sum_{x^n,\hat{s}^n} p(x^n,\hat{s}^n)\frac{1}{n}\sum_{i=1}^{n}\hat{d}_s(x_i,\hat{s}_i)\\
&= \sum_{x^n,\hat{s}^n} p(x^n,\hat{s}^n)\hat{d}_s(x^n,\hat{s}^n)\\
&= \mathbb{E}\left[\hat{d}_s(X^n,\hat{S}^n)\right], \qquad (53)
\end{aligned}
$$

where $\bar{x}_i = (x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n)$, $\bar{s}_i = (s_1,\ldots,s_{i-1},s_{i+1},\ldots,s_n)$, (a) is due to the existence of the Markov chain $S^n \leftrightarrow X^n \leftrightarrow \hat{S}^n$ and hence $p(s^n|x^n) = p(s^n|x^n,\hat{s}^n)$, (b) follows from the definition of block-wise distortion measure in (1), (c) is by the fact that $(S_i,X_i)_{i\in\mathbb{N}}$ is an i.i.d. sequence, and (d) is by the definition of $\hat{d}_s(x,\hat{s})$ in (12). Subsequently, the problem is reduced into a standard lossy source coding problem with two distortion constraints, one on $d_o(x,\hat{x})$ and the other on $\hat{d}_s(x,\hat{s})$. The semantic rate distortion function hence follows from standard achievability and converse proofs [34, Sec. VII] [35, Prob. 7.14] [3, Prob. 10.19].

APPENDIX II
PROOF OF THEOREM 2

The proof of Theorem 2 involves two steps. First we prove that the semantic rate distortion function can be achieved by jointly Gaussian $\hat{X}$ and $\hat{S}$. Then we show that we can further endow a Markov chain structure on X, $\hat{X}$ and $\hat{S}$, so that we only need to optimize with one variable, i.e., $\hat{X}$, while generating $\hat{S}$ from $\hat{X}$ subsequently.

A. Optimality of Jointly Gaussian Reproduction

By the definition of $\hat{d}_s(X;\hat{S})$ in (12), $\mathbb{E}\left[\hat{d}_s(X;\hat{S})\right]$ can be written as follows:

$$
\begin{aligned}
&\mathbb{E}\left[\hat{d}_s(X;\hat{S})\right]\\
&= \int p(x,\hat{s})\left(\int p(s|x)d_s(s,\hat{s})ds\right)dxd\hat{s}\\
&= \int p(x,\hat{s})\\
&\quad\times\left(\int p(\mathbf{H}x+z|x)(\mathbf{H}x+z-\hat{s})(\mathbf{H}x+z-\hat{s})^T dz\right)dxd\hat{s}\\
&\overset{(a)}{=} \int p(x,\hat{s})\Bigg(\int p(z)\,\mathrm{tr}(\mathbf{H}xx^T\mathbf{H}^T+\mathbf{H}xz^T-\mathbf{H}x\hat{s}\\
&\quad + zx^T\mathbf{H}+zz^T-z\hat{s}^T-\hat{s}x^T\mathbf{H}^T-\hat{s}z^T+\hat{s}\hat{s}^T)dz\Bigg)dxd\hat{s}\\
&\overset{(b)}{=} \int p(x,\hat{s})\\
&\quad\times\mathrm{tr}(\mathbf{H}xx^T\mathbf{H}^T-\mathbf{H}x\hat{s}+\mathbf{K}_Z-\hat{s}x^T\mathbf{H}^T+\hat{s}\hat{s}^T)dxd\hat{s}\\
&= \mathrm{tr}(\mathbf{H}\mathbf{K}_X\mathbf{H}^T-\mathbf{H}\mathbf{K}_{X\hat{S}}+\mathbf{K}_Z-\mathbf{K}_{\hat{S}X}\mathbf{H}^T+\mathbf{K}_{\hat{S}})\\
&\overset{(c)}{=} \mathrm{tr}(\mathbf{H}\mathbf{K}_X\mathbf{H}^T-2\mathbf{H}\mathbf{K}_{X\hat{S}}+\mathbf{K}_Z+\mathbf{K}_{\hat{S}}), \qquad (54)
\end{aligned}
$$

where (a) is due to independence between Z and X, (b) is according to the problem setup that $\mathbb{E}(Z)=0$, and (c) is due to the fact that $\mathrm{tr}(\mathbf{H}\mathbf{K}_{X\hat{S}}) = \mathrm{tr}(\mathbf{K}_{\hat{S}X}\mathbf{H}^T)$. From this chain of identities, we see that for any two reproductions of the intrinsic state, $\hat{S}$ and $\hat{S}'$, we have $\mathbb{E}\left[d_s(S;\hat{S})\right] = \mathbb{E}\left[d_s(S;\hat{S}')\right]$ as long as $\mathbf{K}_{\hat{S}} = \mathbf{K}_{\hat{S}'}$ and $\mathbf{K}_{X\hat{S}} = \mathbf{K}_{X\hat{S}'}$.

Therefore, by Theorem 1, the semantic rate distortion function $R_{\mathcal{G}}(D_s,D_o)$ can be further written as

$$
\begin{aligned}
&R_{\mathcal{G}}(D_s,D_o)\\
&\quad= \min I(X;\hat{S},\hat{X}) = h(X) - \max h(X|\hat{S},\hat{X}) \qquad (55)\\
&\qquad\text{s.t. } \mathrm{tr}(\mathbf{K}_X - 2\mathbf{K}_{X\hat{X}} + \mathbf{K}_{\hat{X}}) \le D_o \qquad (56)\\
&\qquad\quad \mathrm{tr}(\mathbf{H}\mathbf{K}_X\mathbf{H}^T - 2\mathbf{H}\mathbf{K}_{X\hat{S}} + \mathbf{K}_Z + \mathbf{K}_{\hat{S}}) \le D_s. \qquad (57)
\end{aligned}
$$

Notice that, by denoting $T \triangleq (\hat{S},\hat{X})$ for convenience, $h(X|\hat{S},\hat{X})$ can be upper bounded as

$$
\begin{aligned}
&h(X|\hat{S},\hat{X})\\
&\quad= h(X|T)\\
&\quad= h(X-\mathbf{K}_{XT}\mathbf{K}_T^{-1}T|T)\\
&\quad\overset{(a)}{\le} h(X-\mathbf{K}_{XT}\mathbf{K}_T^{-1}T)\\
&\quad\overset{(b)}{\le} \frac{1}{2}\log\det(2\pi e\mathbf{K}_{X-\mathbf{K}_{XT}\mathbf{K}_T^{-1}T})\\
&\quad= \frac{1}{2}\log\det(2\pi e(\mathbf{K}_X-\mathbf{K}_{XT}\mathbf{K}_T^{-1}\mathbf{K}_{TX})), \qquad (58)
\end{aligned}
$$

where (a) is by the fact that conditioning reduces entropy, and equality holds when $X-\mathbf{K}_{XT}\mathbf{K}_T^{-1}T$ is independent of T; (b) is due to the fact that Gaussian distribution maximizes differential entropy with given second central moment. Overall, we can see that this upper bound of $h(X|\hat{S},\hat{X})$ is achieved when X and T are jointly Gaussian.

Based on the argument above, for an arbitrary $T=(\hat{S},\hat{X})$, we can generate $T'=(\hat{S}',\hat{X}')$ according to a linear relationship

$$(\hat{S}',\hat{X}') = \mathbf{K}_{TX}\mathbf{K}_X^{-1}X + N, \qquad (59)$$

where N is a multivariate Gaussian random variable following $\mathcal{N}(0,\mathbf{K}_T-\mathbf{K}_{TX}\mathbf{K}_X^{-1}\mathbf{K}_{XT})$ and is independent of X. Clearly it holds that $\mathbf{K}_{T'} = \mathbf{K}_T$ and $\mathbf{K}_{XT} = \mathbf{K}_{XT'}$. According to (58), we can see that $h(X|\hat{S},\hat{X}) \le h(\hat{S}',\hat{X}')$. That is to say, for any $(\hat{S},\hat{X})$ that satisfies the distortion constraints, there always exists a Gaussian $(\hat{S}',\hat{X}')$ which also satisfies the distortion constraints, but achieving a lower code rate. We thus establish that jointly Gaussian reproduction $(\hat{S},\hat{X})$ achieves the semantic rate distortion function.

B. Reduction to One Optimization Variable

In fact, it is unnecessary to optimize with two random variables $(\hat{S},\hat{X})$ simultaneously, and in the following we

reduce the number of optimization variables to only one. We choose the new optimization variable as $\mathrm{cov}(X|\hat{X},\hat{S})$, defined as

$$\mathrm{cov}(X|\hat{X},\hat{S})=\mathbb{E}\left[\left(X-\mathbb{E}\left[X|\hat{X},\hat{S}\right]\right)\left(X-\mathbb{E}\left[X|\hat{X},\hat{S}\right]\right)^T\right],$$

i.e., the error covariance matrix of MMSE estimating X by $(\hat{X},\hat{S})$. By denoting $\mathrm{cov}(X|\hat{X},\hat{S})$ as $\mathbf{\Delta}$ for short, we can write $I(X;\hat{X},\hat{S})$ as (26). Therefore, now the key point is to show that the feasible region defined by (56)-(57) (denoted as $\mathcal{R}_1$) is the same as the feasible region defined by (27)-(29) (denoted as $\mathcal{R}_2$).

First we show that $\mathcal{R}_1 \subseteq \mathcal{R}_2$. For any $\mathbf{K}_{(\hat{S},\hat{X})} \in \mathcal{R}_1$, with $\mathbf{\Delta} = \mathrm{cov}(X|\hat{S},\hat{X})$, we have $\mathbf{\Delta} \preceq \mathrm{cov}(X|\hat{X})$ and $\mathbf{\Delta} \preceq \mathrm{cov}(X|\hat{S})$, and correspondingly $\mathrm{tr}(\mathbf{\Delta}) \leq \mathrm{tr}(\mathrm{cov}(X|\hat{X})) \leq D_o$ and

$$\begin{aligned}\mathrm{tr}(\mathbf{H}\mathbf{\Delta}\mathbf{H}^T+\mathbf{K}_Z) &\leq \mathrm{tr}(\mathbf{H}\mathrm{cov}(X|\hat{S})\mathbf{H}^T+\mathbf{K}_Z)\\ &= \mathrm{tr}(\mathrm{cov}(\mathbf{H}X+Z|\hat{S})) \leq D_s. \end{aligned} \tag{60}$$

That is to say, for any $\mathbf{K}_{(\hat{S},\hat{X})} \in \mathcal{R}_1$, we can find a corresponding $\mathbf{\Delta} \in \mathcal{R}_2$, and hence $\mathcal{R}_1 \subseteq \mathcal{R}_2$.

Then we show that $\mathcal{R}_2 \subseteq \mathcal{R}_1$. For any $\mathbf{\Delta} \in \mathcal{R}_2$, we consider a test channel with $X = \hat{X} + N$ and let $\hat{S} = \mathbf{H}\hat{X}$, where N obeys Gaussian distribution $\mathcal{N}(0,\mathbf{\Delta})$. Hence we have

$$\mathbb{E}\left[d_o(X,\hat{X})\right] = \mathrm{tr}(\mathbf{\Delta}) \leq D_o, \tag{61}$$

$$\begin{aligned}\mathbb{E}\left[d_s(S,\hat{S})\right] &= \mathrm{tr}(\mathbf{H}\mathrm{cov}(X|\hat{S})\mathbf{H}^T+\mathbf{K}_Z)\\ &= \mathrm{tr}(\mathbf{H}\mathbf{\Delta}\mathbf{H}^T+\mathbf{K}_Z) \leq D_s. \end{aligned} \tag{62}$$

That is to say, for any $\mathbf{\Delta} \in \mathcal{R}_2$, we can also find a corresponding tuple of $\mathbf{K}_{(\hat{S},\hat{X})} \in \mathcal{R}_1$, and hence $\mathcal{R}_2 \subseteq \mathcal{R}_1$.

Now, we can conclude that, under the setting of Theorem 2, Theorems 1 and 2 define two optimization problems with the same objective function and the same feasible region. This therefore completes the proof.

Appendix III
Proof of Corollary 4

By Theorem 2 and the identities $\mathbf{H} = \mathbf{K}_{SX}\mathbf{K}_X^{-1}$ and $\mathbf{K}_Z = \mathbf{K}_S - \mathbf{K}_{SX}\mathbf{K}_X^{-1}\mathbf{K}_{SX}^T$ in (21), the semantic rate distortion function of a jointly Gaussian semantic source with covariance matrix (20) is given by

$$R_{\mathcal{G}}(D_s,D_o) = \min_{\mathbf{\Delta}\in\mathcal{S}_m} \frac{1}{2}\log\left(\frac{\det(\mathbf{K}_X)}{\det(\mathbf{\Delta})}\right) \tag{63}$$

$$\text{s.t. } \mathbf{O} \prec \mathbf{\Delta} \preceq \mathbf{K}_X, \tag{64}$$

$$\begin{aligned}&\mathrm{tr}(\mathbf{K}_{SX}\mathbf{K}_X^{-1}\mathbf{\Delta}\mathbf{K}_X^{-1}\mathbf{K}_{SX}^T)\\ &\leq D_s - \mathrm{tr}(\mathbf{K}_S - \mathbf{K}_{SX}\mathbf{K}_X^{-1}\mathbf{K}_{SX}^T),\end{aligned} \tag{65}$$

$$\mathrm{tr}(\mathbf{\Delta}) \leq D_o. \tag{66}$$

We will prove

$$R(D_s,D_o) \leq \frac{1}{2}\log\left(\frac{\det(\mathbf{K}_X)}{\det(\mathbf{\Delta})}\right) \tag{67}$$

for an arbitrary symmetric matrix $\mathbf{\Delta}$ that satisfies (64), (65) and (66), by constructing a test channel. This implies that $R(D_s,D_o)$ is no greater than (63).

In order to construct the test channel, let U be a Gaussian vector with zero mean and covariance matrix $\mathbf{\Delta} - \mathbf{\Delta}\mathbf{K}_X^{-1}\mathbf{\Delta}$, independent of (S,X). That $\mathbf{\Delta} - \mathbf{\Delta}\mathbf{K}_X^{-1}\mathbf{\Delta}$ is semi-definite will be proved in Lemma 2 at the end of this subsection. Define $\hat{X} = (\mathbf{I}_m - \mathbf{\Delta}\mathbf{K}_X^{-1})X + U$ and $\hat{S} = \mathbf{K}_{SX}\mathbf{K}_X^{-1}\hat{X}$. Thus $S \leftrightarrow X \leftrightarrow \hat{X} \leftrightarrow \hat{S}$ is a Markov chain. We will verify in the next paragraphs that $\mathbb{E}[\hat{d}_s(X,\hat{S})] \leq D_s$, where $\hat{d}_s(x,\hat{s}) = \mathbb{E}[\|S-\hat{s}\|_2^2|X=x]$, $\mathbb{E}[\|X-\hat{X}\|_2^2] \leq D_o$, and

$$I(X;\hat{S},\hat{X}) \leq \frac{1}{2}\log\left(\frac{\det(\mathbf{K}_X)}{\det(\mathbf{\Delta})}\right). \tag{68}$$

These leads to (67), and thus proves Corollary 4.

By the definitions of $\hat{X}$ and $\hat{S}$, we have $S-\hat{S} = S - \mathbf{K}_{SX}\mathbf{L}X - \mathbf{K}_{SX}\mathbf{K}_X^{-1}U$, where $\mathbf{L} = \mathbf{K}_X^{-1} - \mathbf{K}_X^{-1}\mathbf{\Delta}\mathbf{K}_X^{-1}$. Noticing $\mathbb{E}[SU^T] = \mathbf{O}_{l\times m}$ and $\mathbb{E}[XU^T] = \mathbf{O}_{m\times m}$, we can obtain, after some algebraic manipulations,

$$\begin{aligned}&\mathbb{E}\left[(S-\hat{S})(S-\hat{S})^T\right]\\ &\quad= \mathbf{K}_S - \mathbf{K}_{SX}\mathbf{K}_X^{-1}\mathbf{K}_{SX}^T + \mathbf{K}_{SX}\mathbf{K}_X^{-1}\mathbf{\Delta}\mathbf{K}_X^{-1}\mathbf{K}_{SX}^T.\end{aligned}$$

Taking the trace in this equation and using (65), we get $\mathbb{E}[\|S-\hat{S}\|_2^2] \leq D_s$. Similar calculations lead to $\mathbb{E}[\|X-\hat{X}\|_2^2] \leq D_o$. For every $x \in \mathbb{R}^m$ and every $\hat{s} \in \mathbb{R}^l$, we have

$$\begin{aligned}\mathbb{E}[\|S-\hat{S}\|_2^2|X=x,\hat{S}=\hat{s}] &= \mathbb{E}[\|S-\hat{s}\|_2^2|X=x,\hat{S}=\hat{s}]\\ &= \mathbb{E}[\|S-\hat{s}\|_2^2|X=x]\\ &= \hat{d}_s(x,\hat{s}),\end{aligned}$$

where the second equality is due to $S \leftrightarrow X \leftrightarrow \hat{S}$. An application of the law of total expectation immediately leads to $\mathbb{E}[\hat{d}_s(X,\hat{S})] = \mathbb{E}[\|S-\hat{S}\|_2^2] \leq D_s$.

It remains to verify (68). We have

$$\begin{aligned}I(X;\hat{S},\hat{X}) &\overset{(a)}{=} I(X;\hat{X})\\ &= h(\hat{X}) - h(\hat{X}|X)\\ &\overset{(b)}{=} h(\hat{X}) - \frac{1}{2}\log((2\pi e)^m\det(\mathbf{\Delta}-\mathbf{\Delta}\mathbf{K}_X^{-1}\mathbf{\Delta}))\\ &\overset{(c)}{\leq} \frac{1}{2}\log((2\pi e)^m(\mathbf{K}_X-\mathbf{\Delta}))\\ &\quad - \frac{1}{2}\log((2\pi e)^m\det(\mathbf{\Delta}-\mathbf{\Delta}\mathbf{K}_X^{-1}\mathbf{\Delta}))\\ &= \frac{1}{2}\log\left(\frac{\det(\mathbf{K}_X)}{\det(\mathbf{\Delta})}\right),\end{aligned}$$

where (a) is by $X \leftrightarrow \hat{X} \leftrightarrow \hat{S}$, (b) is because after translation $h(\hat{X}|X) = h(U)$, and (c) is because the Gaussian distribution maximizes the differential entropy subject to a covariance constraint.

Finally let us verify the existence of the auxiliary random vector U.

Lemma 2: For any $\mathbf{\Delta}$, $\mathbf{K} \in \mathcal{S}_m$, $\mathbf{\Delta} \preceq \mathbf{K}$, $\mathbf{\Delta} - \mathbf{\Delta}\mathbf{K}^{-1}\mathbf{\Delta}$ is semi-definite.

Proof: Because $\mathbf{\Delta}$ is positive definite, there exists an $m\times m$ matrix $\mathbf{Q}$ such that $\mathbf{\Delta} = \mathbf{Q}^T\mathbf{Q}$. For every $\lambda < 0$,

$$\begin{aligned}&\det(\lambda\mathbf{I}_m - (\mathbf{I}_m - \mathbf{Q}\mathbf{K}^{-1}\mathbf{Q}^T))\\ &\quad= (\lambda-1)^m\det\left(\mathbf{I}_m + \frac{1}{\lambda-1}\mathbf{Q}\mathbf{K}^{-1}\mathbf{Q}^T\right)\end{aligned}$$

$$
\begin{aligned}
&= (\lambda-1)^m \det\left(\mathbf{I}_m + \frac{1}{\lambda-1}\mathbf{K}^{-1}\mathbf{Q}^T\mathbf{Q}\right)\\
&= (\lambda-1)^m \det\left(\mathbf{K}^{-1}\left(\mathbf{K} + \frac{1}{\lambda-1}\mathbf{\Delta}\right)\right)\\
&= \frac{\det((\lambda-1)\mathbf{K}+\mathbf{\Delta})}{\det(\mathbf{K})} \neq 0,
\end{aligned}
$$

because $(\lambda-1)\mathbf{K}+\mathbf{\Delta} = \lambda\mathbf{K} - (\mathbf{K}-\mathbf{\Delta})$ is negative definite. So $\mathbf{I}_m - \mathbf{Q}\mathbf{K}^{-1}\mathbf{Q}^T$ does not have any negative eigenvalue. Therefore $\mathbf{I}_m - \mathbf{Q}\mathbf{K}^{-1}\mathbf{Q}^T$ is positive semi-definite, and consequently $\mathbf{\Delta} - \mathbf{\Delta}\mathbf{K}^{-1}\mathbf{\Delta} = \mathbf{Q}^T(\mathbf{I}_m - \mathbf{Q}\mathbf{K}^{-1}\mathbf{Q}^T)\mathbf{Q}$ is also positive semi-definite. □

APPENDIX IV
DERIVATION OF THE WEIGHTED REVERSE WATER-FILLING SOLUTION

We first rewrite (26) with a variable substitution $\mathbf{\Delta} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\dagger$. This leads to

$$
R_{\mathcal{G}}(D_s, D_o) = \min_{\mathbf{D}\in B(D_s,D_o)} \frac{1}{2}\log\left(\frac{\det(\mathbf{K}_X)}{\det(\mathbf{D})}\right),
$$

where $B(D_s, D_o)$ is the set of positive definite real matrices $\mathbf{D}$ that satisfy

$$
\begin{aligned}
\mathbf{D} &\preceq \mathbf{Q}^\dagger\mathbf{K}_X\mathbf{Q},\\
\operatorname{tr}(\mathbf{Q}^\dagger\mathbf{H}^T\mathbf{H}\mathbf{Q}\mathbf{D}) &\leq D_s - \operatorname{tr}(\mathbf{K}_Z),\\
\operatorname{tr}(\mathbf{D}) &\leq D_o.
\end{aligned}
$$

Any optimal $\mathbf{D}$ in this minimization is diagonal. To see this, consider a non-diagonal $\mathbf{D} \in B(D_s, D_o)$. Replacing the non-diagonal elements in $\mathbf{D}$ with zeros, we get a new matrix $\mathbf{D}' = \operatorname{diag}(\delta_1, \delta_2, \cdots, \delta_m)$. Because

$$
\mathbf{O}_m \prec \mathbf{D} \preceq \mathbf{Q}^\dagger\mathbf{K}_X\mathbf{Q} = \operatorname{diag}(\sigma_1, \sigma_2, \cdots, \sigma_m),
$$

we have $0 < \delta_j \leq \sigma_j$ for each $j \in \{1, 2, \cdots, m\}$, which impies $\mathbf{O}_m \prec \mathbf{D}' \preceq \mathbf{Q}^\dagger\mathbf{K}_X\mathbf{Q}$. Moreover,

$$
\begin{aligned}
\operatorname{tr}(\mathbf{Q}^\dagger\mathbf{H}^T\mathbf{H}\mathbf{Q}\mathbf{D}') &= \sum_{j=1}^m \alpha_j\delta_j = \operatorname{tr}(\mathbf{Q}^\dagger\mathbf{H}^T\mathbf{H}\mathbf{Q}\mathbf{D})\\
&\leq D_s - \operatorname{tr}(\mathbf{K}_Z),\\
\operatorname{tr}(\mathbf{D}') &= \sum_{j=1}^m \delta_j = \operatorname{tr}(\mathbf{D})\\
&\leq D_o.
\end{aligned}
$$

So $\mathbf{D}' \in B(D_s, D_o)$. By Hadamard's inequality,

$$
\frac{1}{2}\log\left(\frac{\det(\mathbf{K}_X)}{\det(\mathbf{D}')}\right) < \frac{1}{2}\log\left(\frac{\det(\mathbf{K}_X)}{\det(\mathbf{D})}\right).
$$

Therefore, any non-diagonal $\mathbf{D} \in B(D_s, D_o)$ is suboptimal, and (43) is verified.

By the Karush-Kuhn-Tucker (KKT) optimality conditions, there exist non-negative numbers λ, μ, ν_1, ν_2, $\cdots$, ν_m that satisfy

$$
\lambda\left(\sum_{j=1}^m \delta_j^* - D_o\right) = 0,
$$

$$
\begin{aligned}
\mu\left(\sum_{j=1}^m \alpha_j\delta_j^* - D_s + \operatorname{tr}(\mathbf{K}_Z)\right) &= 0,\\
\nu_j(\delta_j^* - \sigma_j) &= 0, \quad \forall j \in \{1, 2, \cdots, m\},\\
-\frac{1}{\delta_j^*} + \lambda + \mu\alpha_j + \nu_j &= 0, \quad \forall j \in \{1, 2, \cdots, m\}.
\end{aligned}
$$

Suppose $\lambda = 0$ and $\mu = 0$. For each $j \in \{1, 2, \cdots, m\}$, we have $\nu_j = 1/\delta_j^* > 0$, so $\delta_j^* = \sigma_j$. Because $\delta_1^*, \delta_2^*, \cdots, \delta_m^*$ satisfy (45) and (46), we have

$$
\begin{aligned}
D_s &\geq \sum_{j=1}^m \alpha_j\sigma_j + \operatorname{tr}(\mathbf{K}_Z) = \operatorname{tr}(\mathbf{H}\mathbf{K}_X\mathbf{H}^T + \mathbf{K}_Z),\\
D_o &\geq \sum_{j=1}^m \sigma_j = \operatorname{tr}(\mathbf{K}_X),
\end{aligned}
$$

i.e. $(D_s, D_o) \in A_0$.

Suppose $\lambda > 0$ and $\mu = 0$. The problem now reduces to the one involved in the rate distortion problem of parallel Gaussian sources [3], because the constraint (46) is active and (45) is not. Thus (50) holds, and

$$
\begin{aligned}
D_o &= \sum_{j=1}^m \delta_j^* = \sum_{j=1}^m \min\left(\sigma_j, \frac{1}{\lambda}\right),\\
D_s &\geq \sum_{j=1}^m \alpha_j\delta_j^* + \operatorname{tr}(\mathbf{K}_Z) = \sum_{j=1}^m \alpha_j \min\left(\sigma_j, \frac{1}{\lambda}\right) + \operatorname{tr}(\mathbf{K}_Z),
\end{aligned}
$$

i.e. $(D_s, D_o) \in A_1$.

Similarly, the conditions $\lambda = 0$ and $\mu > 0$ imply (51) leading to $(D_s, D_o) \in A_2$, and the conditions $\lambda > 0$ and $\mu > 0$ imply (52) leading to $(D_s, D_o) \in A_3$.

REFERENCES

[1] J. Liu, W. Zhang, and H. V. Poor, "A rate-distortion framework for characterizing semantic information," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2021, pp. 2894–2899.

[2] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[4] S. Ma, X. Zhang, S. Wang, X. Zhang, C. Jia, and S. Wang, "Joint feature and texture coding: Toward smart video representation via front-end intelligence," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 3095–3105, Oct. 2019.

[5] L. Duan, J. Liu, W. Yang, T. Huang, and W. Gao, "Video coding for machines: A paradigm of collaborative compression and intelligent analytics," *IEEE Trans. Image Process.*, vol. 29, pp. 8680–8695, 2020.

[6] S. Yang, Y. Hu, W. Yang, L.-Y. Duan, and J. Liu, "Towards coding for human and machine vision: Scalable face image coding," *IEEE Trans. Multimedia*, vol. 23, pp. 2957–2971, 2021.

[7] Y. Yang, G. Shu, and M. Shah, "Semi-supervised learning of feature hierarchies for object detection in a video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1650–1657.

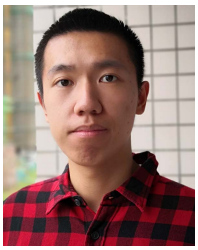
[8] Y. Wu *et al.*, "Person reidentification by multiscale feature representation learning with random batch feature mask," *IEEE Trans. Cognit. Develop. Syst.*, vol. 13, no. 4, pp. 865–874, Dec. 2021.

[9] K. Liu, D. Liu, L. Li, N. Yan, and H. Li, "Semantics-to-signal scalable image compression with learned revertible representations," *Int. J. Comput. Vis.*, vol. 129, no. 9, pp. 2605–2621, Jun. 2021.

[10] L. R. Rabiner and R. W. Schafer, "Introduction to digital speech processing," *Found. Trends Signal Process.*, vol. 1, nos. 1–2, pp. 1–194, 2007.

[11] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 2, pp. 254–272, Apr. 1981.

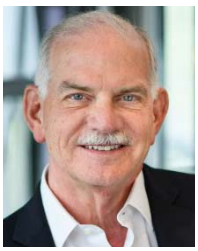
[12] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE Nat. Conv. Rec.*, vol. 4, pp. 142–163, Mar. 1959.
[13] P. Popovski, O. Simeone, F. Boccardi, D. Gunduz, and O. Sahin, "Semantic-effectiveness filtering and control for post-5G wireless connectivity," 2019, *arXiv:1907.02441*.
[14] M. Kountouris and N. Pappas, "Semantics-empowered communication for networked intelligent systems," 2020, *arXiv:2007.11579*.
[15] H. Seo, J. Park, M. Bennis, and M. Debbah, "Semantics-native communication with contextual reasoning," 2021, *arXiv:2108.05681*.
[16] Y. Bar-Hillel and R. Carnap, "Semantic information," *Brit. J. Philosophy Sci.*, vol. 4, no. 14, pp. 147–157, 1953.
[17] L. Floridi, "Outline of a theory of strongly semantic information," *Minds Mach.*, vol. 14, no. 2, pp. 197–221, May 2004.
[18] J. Bao *et al.*, "Towards a theory of semantic communication," in *Proc. IEEE Netw. Sci. Workshop*, Jun. 2011, pp. 110–117.
[19] B. Juba, *Universal Semantic Communication*. Berlin, Germany: Springer, 2011.
[20] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Annu. Allerton Conf. Commun. Control Comput.*, Monticello, IL, USA, Sep. 1999, pp. 368–377.
[21] Z. Goldfeld and Y. Polyanskiy, "The information bottleneck problem and its applications in machine learning," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 19–38, May 2020.
[22] A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Medard, "From the information bottleneck to the privacy funnel," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Nov. 2014, pp. 501–505.
[23] Y. Y. Shkel, R. S. Blum, and H. V. Poor, "Secrecy by design with applications to privacy and compression," *IEEE Trans. Inf. Theory*, vol. 67, no. 2, pp. 824–843, Feb. 2021.
[24] N. Shlezinger, Y. C. Eldar, and M. R. D. Rodrigues, "Hardware-limited task-based quantization," *IEEE Trans. Signal Process.*, vol. 67, no. 20, pp. 5223–5238, Oct. 2019.
[25] Y. Blau and T. Michaeli, "Rethinking lossy compression: The rate-distortion-perception tradeoff," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jun. 2019, pp. 675–685.
[26] A. Kipnis, S. Rini, and A. J. Goldsmith, "The rate-distortion risk in estimation from compressed data," *IEEE Trans. Inf. Theory*, vol. 67, no. 5, pp. 2910–2924, May 2021.
[27] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.
[28] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2011.
[29] R. W. Yeung, *Information Theory and Network Coding* (Information Technology: Transmission, Processing and Storage). New York, NY, USA: Springer, 2008.
[30] R. Dobrushin and B. Tsybakov, "Information transmission with additional noise," *IRE Trans. Inf. Theory*, vol. 8, no. 5, pp. 293–304, Sep. 1962.
[31] J. K. Wolf and J. Ziv, "Transmission of noisy information to a noisy receiver with minimum distortion," *IEEE Trans. Inf. Theory*, vol. IT-16, no. 4, pp. 406–411, Jul. 1970.
[32] T. Berger, *Rate Distortion Theory*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1971.
[33] H. S. Witsenhausen, "Indirect rate distortion problems," *IEEE Trans. Inf. Theory*, vol. IT-26, no. 5, pp. 518–521, Sep. 1980.
[34] A. A. El Gamal and T. M. Cover, "Achievable rates for multiple descriptions," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 6, pp. 851–857, Nov. 1982.
[35] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2011.
[36] Y. Xia, C. Sun, and W. X. Zheng, "Discrete-time neural network for fast solving large linear L_1 estimation problems and its application to image restoration," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 5, pp. 812–820, Mar. 2012.
[37] Y. Oohama, "The rate-distortion function for the quadratic Gaussian CEO problem," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1057–1070, May 1998.
[38] R. M. Gray, *Toeplitz and Circulant Matrices: A Review*. Boston, MA, USA: NOW, 2009.



Jiakun Liu received the B.S. degree in electronic information engineering from the University of Science and Technology of China, Hefei, China, in 2018, where he is currently pursuing the Ph.D. degree. His research focuses on information theory and statistical learning.

Shuo Shao (Member, IEEE) received the B.S. degree in information science from Southeast University, China, in 2011, the M.A.Sc. degree in electrical and computer engineering from McMaster University, Canada, in 2013, and the Ph.D. degree from Texas A&M University, USA, in 2017. Since 2017, he has been with the School of Electronics, Information, and Electrical Engineering, Shanghai Jiao Tong University, China. His research interests are in network information theory, algebraic code, and machine learning.

Wenyi Zhang (Senior Member, IEEE) received the bachelor's degree in automation from Tsinghua University, Beijing, China, in 2001, and the master's and Ph.D. degrees in electrical engineering from the University of Notre Dame, Notre Dame, IN, USA, in 2003 and 2006, respectively. He was with the Communication Science Institute, University of Southern California, as a Post-Doctoral Research Associate and with Qualcomm Inc., Corporate Research and Development. He is currently a Professor with the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China. His research interests include wireless communications and networking, information theory, and statistical signal processing. He was an Editor of IEEE COMMUNICATIONS LETTERS and IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.



H. Vincent Poor (Life Fellow, IEEE) received the Ph.D. degree in EECS from Princeton University in 1977. From 1977 to 1990, he was on the faculty of the University of Illinois at Urbana–Champaign. Since 1990, he has been on the faculty of Princeton University, where he is currently the Michael Henry Strater University Professor. During 2006 to 2016, he served as the Dean of Princeton's School of Engineering and Applied Science. He has also held visiting appointments at several other universities, including most recently at Berkeley and Cambridge. His research interests are in the areas of information theory, machine learning and network science and their applications in wireless networks, and energy systems and related fields. Among his publications in these areas is the forthcoming book *Machine Learning and Wireless Communications* (Cambridge University Press). He is a member of the National Academy of Engineering and the National Academy of Sciences and is a foreign member of the Chinese Academy of Sciences, the Royal Society, and other national and international academies. He received the IEEE Alexander Graham Bell Medal in 2017.