On Information Bottleneck for Gaussian Processes

Michael Dikshtein, Nir Weinberger, and Shlomo Shamai (Shitz)
Department of Electrical and Computer Engineering, Technion, Haifa 3200003, Israel
Email: {michaeldic@campus., nirwein@, sshlomo@ee.}technion.ac.il

2022 IEEE Information Theory Workshop (ITW) | 978-1-6654-8341-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/ITW54588.2022.9965939

***Abstract*—The *information bottleneck* (IB) problem of jointly stationary Gaussian sources is considered. A water-filling solution for the IB rate is given in terms of its SNR spectrum and whose rate is attained via frequency domain test-channel realization. A time-domain realization of the IB rate, based on linear prediction, is also proposed, which lends itself to an efficient implementation of the corresponding remote source-coding problem. A compound version of the problem is addressed, in which the joint distribution of the source is not precisely specified but rather in terms of a lower bound on the guaranteed mutual information. It is proved that a white SNR spectrum is optimal for this setting.**

I. Introduction

It is by now well established that the *information bottleneck* (IB) method [1] plays a central role in information theory, machine learning, and various other fields. In its basic form, it introduced an alternative conceptual approach to the problem of lossy compression. In classic rate-distortion problems [2], a distortion measure must be provided in order to calculate the respective rate-distortion function and characterize the single-letter compression strategy. However, in practice, determining the distortion measure may be a challenging problem on its own. One motivating example is speech compression, in which it is difficult to identify the prominent features of the vocal signal that affect the quality of the reconstructed audio. Also, one may argue that the transcript of the speech signal has more information than the features of the analog waveform. The IB framework tackles those issues by introducing an additional, informative variable to the problem that acts as the appropriate labeling of the data to be compressed. Thus, IB provides a natural fidelity measure in cases where such cannot be defined or does not exist; an essential feature for modern source coding problems. We also mention that, as is well known, the IB is the information-theoretic solution to a remote source coding problem [3], [4], when the log-loss is chosen as the distortion measure. Nonetheless, the setting behind remote-source coding is fundamentally different from IB, despite the formal similarity between the resulting optimization problems.

In general, the IB function is a non-convex optimization problem and does not have a closed-form solution. However, it can be analytically solved in some canonical scenarios, mainly the *doubly symmetric binary source* (DSBS) in [5],

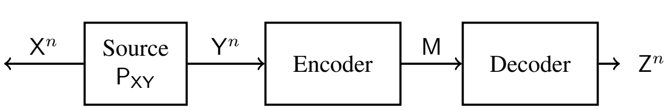

Fig. 1: Block diagram of Remote Source Coding.

and a jointly Gaussian vector source in [6]. In all other cases, as proposed in [1], a Blahut-Arimoto type alternating minimization algorithm [7], [8], is typically used for obtaining a solution. This algorithm was shown to converge to a local stationary point but has no global convergence guarantees.

This work addresses the bivariate Gaussian setting and replaces the finite length vectors formulation with that of a stationary Gaussian random process. Considering random processes instead of finite length vectors gives the problem a more natural and practical flavor and is motivated by the communication and signal-processing problems, in which the receiver sequentially processes the samples. It has also rooted in remote source coding and signal-denoising applications. This wide-sense stationary setting is usually approached via frequency-domain methods, leading to linear time-invariant system implementation. In this paper, we will also propose a prediction-based scheme that implements the compression phase using a time-domain single-letter, sequential processing.

A. Problem Formulation

Consider the discrete-time IB model illustrated in Fig. 1. The real valued bivariate source

$$(\{\mathsf{X}_t\}, \{\mathsf{Y}_t\}) = \dots, (\mathsf{X}_{-1}, \mathsf{Y}_{-1}), (\mathsf{X}_0, \mathsf{Y}_0), (\mathsf{X}_1, \mathsf{Y}_1), \dots, \quad (1)$$

is a bivariate stationary Gaussian random process, with marginal power spectral densities $S_{\mathsf{X}}(f)$, $S_{\mathsf{Y}}(f)$, and cross-power spectrum $S_{\mathsf{XY}}(f)$. The encoder observes Y^n and maps it to a compressed representation with an index $\mathsf{M} \in [1 : 2^{nC}]$. The decoder recovers Z^n from M. The compression strategy $\mathsf{P}_{\mathsf{Z}|\mathsf{Y}}$ is optimized to maximize the normalized mutual information between X^n and Z^n.

The considered jointly Gaussian bivariate stationary source can be equivalently represented by linear time invariant filters [9, Thm. 4.5.5], i.e., $\mathsf{Y}_n = h_n * \mathsf{X}_n + \mathsf{W}_n$; where $*$ stands for convolution operator, h_n is the impulse response of the linear system with transfer function

$$H(f) = \frac{S_{\mathsf{XY}}(f)}{S_{\mathsf{X}}(f)}, \quad (2)$$

and W_n is an additive colored Gaussian noise with power spectrum

$$S_{\mathsf{W}}(f) = S_{\mathsf{Y}}(f) - \frac{|S_{\mathsf{XY}}(f)|^2}{S_{\mathsf{X}}(f)}. \quad (3)$$

The *information bottleneck* (IB) rate of a bivariate stationary source with memory, which is given as a limit of normalized mutual information associated with vectors of source samples, can be written as

$$\mathcal{R}^{ib}(C) = \lim_{n\to\infty} R_n^{ib}(C; \mathsf{P}_{\mathsf{X}^n\mathsf{Y}^n}), \tag{4}$$

where

$$\begin{aligned} R_n^{ib}(C; \mathsf{P}_{\mathsf{X}^n\mathsf{Y}^n}) = \quad & \sup_{\mathsf{P}_{\mathsf{Z}^n|\mathsf{Y}^n}} && \frac{1}{n} I(\mathsf{X}^n; \mathsf{Z}^n) \\ & \text{subject to} && \frac{1}{n} I(\mathsf{Y}^n; \mathsf{Z}^n) \le C \end{aligned} \tag{5}$$

is the normalized IB function for random vectors $(\mathsf{X}^n, \mathsf{Y}^n, \mathsf{Z}^n)$. We will usually abbreviate $R_n^{ib}(C; \mathsf{P}_{\mathsf{X}^n\mathsf{Y}^n})$ as $R_n^{ib}(C)$. The channel $\mathsf{P}_{\mathsf{Z}^n|\mathsf{Y}^n}$ which achieves this supremum subject to the bottleneck constraint is termed an *optimal test channel*.

Contributions: Before understanding complex models such as DNN, one must first understand the canonical models, such as the Gaussian one. This is also a common theoretical approach in machine learning. For example, there are many works on two-layer neural networks, or even linear networks (which are in some sense trivial since a concatenation of linear operations is just a linear operation on its own). The IB method provides a natural fidelity measure for modern source coding applications. The standard rate-distortion problem for Gaussian processes with least-squares loss has been considered in [14]. Thus, our problem extends previous works by considering a more general distortion measure. Furthermore, stochastic colored process stands as a good model for a general source with continuous stream of output symbols, i.e. audio, video, sensor output. Gaussian processes are a simple model for such processes which are usually manageable for analytical consideration.

We first state a "water filling" solution for the IB function of a stationary bivariate Gaussian source in terms of its SNR spectrum. Since IB is essentially a remote source coding with log-loss distortion [3], [4], the resulting formula is similar in spirit to the capacity formula of power-constrained inter-symbol interference (ISI) channel with Gaussian noise [10], [11], and the rate-distortion function of a stationary Gaussian source [9]. The result above is derived from asymptotic quantities of mutual information between Gaussian random vectors in the frequency domain [12]. We next translate these vector mutual information measures to scalar ones via linear prediction. This parallels a result which states that the capacity of the ISI channel is equal to the single letter mutual information over a slicer embedded in a decision-feedback noise-prediction loop [13], and similarly, that the rate-distortion function equals the single letter mutual information over an *additive white Gaussian noise* (AWGN) channel embedded in a source prediction loop [14]. We show that a parallel result holds for the IB rate $\mathcal{R}^{ib}(C)$, which is equal to the scalar mutual information over an AWGN channel embedded in a source prediction loop, as shown in Fig. 3. This result implies that $\mathcal{R}^{ib}(C)$ can essentially be realized sequentially. Finally, we consider a compound version of the IB problem with Gaussian processes, in which the cross-spectrum of the processes $(\{\mathsf{X}_t\}, \{\mathsf{Y}_t\})$ is not fully specified, and it is only known that the mutual information rate $\mathcal{I}(\{\mathsf{X}_t\}; \{\mathsf{Y}_t\}) \ge C_1$ for some C_1. In general, this problem is motivated by the uncertainty of the correlation between $\{\mathsf{X}_t\}$ and $\{\mathsf{Y}_t\}$, and we refer the reader to [15] for a more elaborate discussion. Here we provide an explicit solution for this scenario.

Omitted proofs and other details are in the full version of this paper [16].

B. Related Work

The vector Gaussian IB setting was first considered in [6]. It was shown that jointly Gaussian vectors triple $\mathbf{X} \to \mathbf{Y} \to \mathbf{Z}$ is optimal [17], and a closed-form formula for the IB curve was obtained. Linear operations can typically obtain optimal solutions for the Gaussian setting. However, channel output compression subject to squared-error distortion was shown to be sub-optimal to the Gaussian IB setting [18]. The latter issue was solved in [19] by adding a prefilter prior the rate-distortion block. The discrete-time Gaussian process setting was considered in [19] and [20].

For the jointly stationary bivariate source $(\{\mathsf{X}_t\}, \{\mathsf{Y}_t\})$, the *Privacy Funnel* (PF) rate [21] is defined by

$$\mathcal{R}^{pf}(C) = \lim_{n\to\infty} R_n^{pf}(C; \mathsf{P}_{\mathsf{Y}^n\mathsf{Z}^n}), \tag{6}$$

where

$$\begin{aligned} R_n^{pf}(C; \mathsf{P}_{\mathsf{Y}^n\mathsf{Z}^n}) = \quad & \inf_{\mathsf{P}_{\mathsf{Z}^n|\mathsf{Y}^n}} && \frac{1}{n} I(\mathsf{X}^n; \mathsf{Z}^n) \\ & \text{subject to} && \frac{1}{n} I(\mathsf{X}^n; \mathsf{Y}^n) \ge C \end{aligned} \tag{7}$$

is the normalized PF function for random vectors $(\mathsf{X}^n, \mathsf{Y}^n, \mathsf{Z}^n)$.

II. Information Bottleneck for Gaussian Processes

In this section, we review and state results on IB for jointly Gaussian random vectors and jointly Gaussian random processes.

A. Review: IB for jointly Gaussian pair $(\mathbf{X}, \mathbf{Y})$

We begin with a brief review of the basic setting in which $(\mathbf{X}, \mathbf{Y})$ in (5) are jointly Gaussian random vectors. Let $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{X}})$, $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{Y}})$ with cross-covariance matrix $\Sigma_{\mathbf{XY}}$. Those bivariate random Gaussian vectors can equivalently represented using a linear additive-noise form, i.e.,

$$\mathbf{Y} = H\mathbf{X} + \mathbf{W}, \tag{8}$$

where $H = \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{X}}^{-1}$ and $\Sigma_{\mathbf{W}} = \Sigma_{\mathbf{Y}} - \Sigma_{\mathbf{XY}}^T \Sigma_{\mathbf{X}}^{-1} \Sigma_{\mathbf{XY}}$. Let $O\Gamma O^T$ be the Singular Value Decomposition (SVD) of the Signal-to-Noise Ratio (SNR) covariance matrix Σ_{snr}, defined by

$$\Sigma_{snr} \triangleq \Sigma_{\mathbf{W}}^{-1/2} H \Sigma_{\mathbf{X}} H^T \Sigma_{\mathbf{W}}^{-1/2}, \tag{9}$$

with $\Gamma = \text{diag}\{\gamma_i\}_{i=1}^n$. The vector Gaussian IB function was first determined in [6] (see also [19] and [22, Sec. 2], where the trade-off parameter β is replaced here by $1+1/\theta$). We next present this result with a structure that resembles the classical rate-distortion function for Gaussian sources with memory [9, Thm. 4.5.3], where θ plays the role of the water-filling level – First θ is chosen to satisfy the bottleneck constraint C, and then the information rate is calculated. The importance of this form of presentation is that it will be used in the next section to obtain realizations in the time domain.

Lemma 1: The normalized IB rate (5) is achieved by a jointly Gaussian triple $(\mathbf{X}, \mathbf{Y}, \mathsf{Z})$ and is given by

$$R_n^{ib}(C) = \frac{1}{2n}\sum_{i=1}^{n}\log\left[\frac{1+\gamma_i}{1+\theta}\right]^+, \tag{10}$$

where θ is the water filling level chosen such that

$$\frac{1}{2n}\sum_{i=1}^{N}\log\left[\frac{\gamma_i}{\theta}\right]^+ = C, \tag{11}$$

with $[x]^+ \triangleq \max\{1, x\}$, and $\{\gamma_i\}_{i=1}^n$ are the eigenvalues of the SNR covariance matrix Σ_{snr}.

B. IB for jointly Gaussian Processes

Equipped with the form of the IB function in Lemma 1, the solution to the discrete-time processes setting follows from Szegö's theorem [12, Thm. 4.2]. While a related result with filtered observation was first published in [19], in our approach such assumption is not required.

Theorem 1: The IB rate (4) for Gaussian random processes is given by

$$\mathcal{R}^{ib}(C) = \frac{1}{2}\int_{-1/2}^{1/2}\log\left[\frac{1+\Gamma(f)}{1+\theta}\right]^+ \mathrm{d}f, \tag{12}$$

where we choose the *water level* θ so that the total rate is C,

$$C = \frac{1}{2}\int_{-1/2}^{1/2}\log\left[\frac{\Gamma(f)}{\theta}\right]^+ \mathrm{d}f, \tag{13}$$

and $\Gamma(f)$ is the SNR spectrum, defined by

$$\Gamma(f) = \frac{|H(f)|^2 S_{\mathsf{X}}(f)}{S_{\mathsf{W}}(f)} = \frac{|S_{\mathsf{XY}}(f)|^2}{S_{\mathsf{X}}(f)S_{\mathsf{Y}}(f) - |S_{\mathsf{XY}}(f)|^2}. \tag{14}$$

We may define the *distortion spectrum* [9] as

$$D_\theta(f) = \begin{cases}\theta, & \Gamma(f) > \theta\\ \Gamma(f), & \text{otherwise}\end{cases} \tag{15}$$

and then Thm. 1 can be equivalently stated as

$$\mathcal{R}^{ib}(C) = \frac{1}{2}\int_{-1/2}^{1/2}\log\left[\frac{1+\Gamma(f)}{1+D_\theta(f)}\right] \mathrm{d}f, \tag{16}$$

and

$$C = \frac{1}{2}\int_{-1/2}^{1/2}\log\left[\frac{\Gamma(f)}{D_\theta(f)}\right] \mathrm{d}f. \tag{17}$$

As immediate consequence of Thm. 1 is that the optimal channel from Y_n to Z_n can be described in a linear AWGN form

$$\mathsf{Z}_n = h_{2,n} * (h_{1,n} * g_n * \omega_n * \mathsf{Y}_n + \mathsf{N}_n), \tag{18}$$

where ω_n, g_n, $h_{1,n}$ and $h_{2,n}$ are impulse responses of a noise-whitening filter, shaping filter, a suitable prefilter and postfilter respectively, whose absolute squared value frequency responses are given by

$$|\Omega(f)|^2 = \frac{1}{S_{\mathsf{W}}(f)}, \quad |G(f)|^2 = \frac{\Gamma(f)}{1+\Gamma(f)}, \tag{19}$$

$$|H_1(f)|^2 = 1 - \frac{D(f)}{S_{\mathsf{Y}}(f)}, \quad H_2(f) = H_1^*(f), \tag{20}$$

and $\mathsf{N}_n \sim \mathcal{N}(0, \theta)$. The respective forward channel realization is illustrated in Fig. 2.

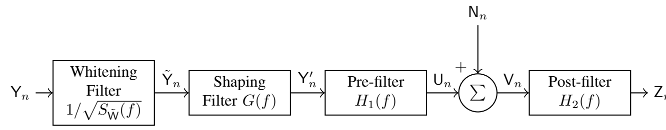

Fig. 2: Forward channel realization.

Remark 1: The absolute frequency response of the shaping filter $G(f)$ is exactly the square root of the noncausal Wiener filter. Thus, the shaping filter plays the role of denoiser in the spirit of MMSE estimation of $\{\mathsf{X}_t\}$ from $\{\mathsf{Y}_t\}$.

III. Remote Source Coding in Time Domain via Prediction

In this section, we consider the compression system illustrated in Fig. 3. This setting is motivated by the sequential structure of *differential pulse-code modulation* (DPCM) [23], which was initially proposed in order to compress highly correlated sequences efficiently. The idea behind DPCM is to translate the encoding of dependent source samples into a series of independent encodings. The latter is accomplished by linear prediction. The input source sample is predicted from previously encoded samples at each instant. The prediction error is encoded by a scalar quantizer and added to the predicted value to form the new reconstruction. The prefilter output, denoted U_n, is fed to the central block, which generates a process V_n according to the following recursion equations:

$$\hat{\mathsf{U}}_n = g(\mathsf{V}_{n-L}^{n-1}) \tag{21}$$

$$\mathsf{W}_n = \mathsf{U}_n - \hat{\mathsf{U}}_n \tag{22}$$

$$\mathsf{Q}_n = \mathsf{W}_n + \mathsf{N}_n \tag{23}$$

$$\mathsf{V}_n = \hat{\mathsf{U}}_n + \mathsf{Q}_n, \tag{24}$$

where $\mathsf{N}_n \sim \mathcal{N}(0, \theta)$ is a zero-mean white Gaussian noise, independent of the input process $\{\mathsf{U}_t\}$, whose variance is equal to the water level θ; and $g(\cdot)$ is some prediction function for the input U_n given the L past samples of the output process (V_{n-L}^{n-1}).

The block from U_n to V_n is equivalent to the configuration of DPCM, with the DPCM quantizer replaced by the additive Gaussian noise channel $\mathsf{Q}_n = \mathsf{W}_n + \mathsf{N}_n$. In particular, the recursive structure implies that this block satisfies the well-known "DPCM error identity":

$$\mathsf{V}_n = \mathsf{U}_n + (\mathsf{Q}_n - \mathsf{W}_n) = \mathsf{U}_n + \mathsf{N}_n. \tag{25}$$

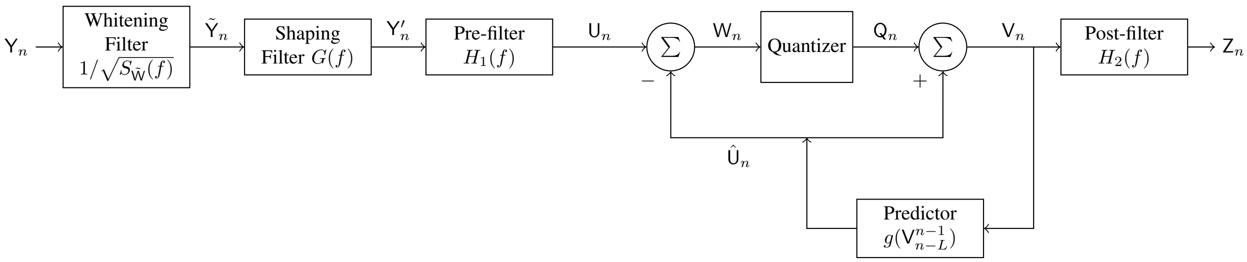

Fig. 3: A predictive quantization scheme.

Namely, the output V_n is a noisy version of U_n passed through the AWGN channel $\mathsf{V}_n = \mathsf{W}_n + \mathsf{N}_n$. Thus, the system of Fig. 3 is equivalent to the forward channel implementation of the Gaussian IB illustrated in Fig. 2.

The prediction function $g(\cdot)$ is chosen to be linear as in DPCM, i.e.,

$$g(\mathsf{V}_{n-L}^{n-1}) = \sum_{i=1}^{L} a_i \mathsf{V}_{n-i}, \tag{26}$$

where $\{a_i\}_{i=1}^{L}$ are chosen to minimize the mean-squared prediction error

$$\sigma_L^2 = \min_{\{a_i\}_{i=1}^{n}} \mathbb{E}\left[\left[\mathsf{U}_n - \sum_{i=1}^{L} a_i \mathsf{V}_{n-i}\right]^2\right]. \tag{27}$$

Since V_n is noisy version of U_n, $g(\cdot)$ is termed as a "noisy predictor". If $\{\mathsf{U}_t\}$ and $\{\mathsf{V}_t\}$ are jointly Gaussian, then the optimal predictor of any order is linear, thus, σ_L^2 is also the MMSE in estimating U_n from V_{n-L}^{n-1}. Clearly, this MMSE is nonincreasing with the prediction depth L, and it converges as L goes to infinity to $\sigma_\infty^2 = \lim_{L\to\infty} \sigma_L^2$, which is the optimal predictor of U_n based on all past observations of V_n^-.

Remark 2: Note that σ_∞^2 is closely related to the entropy power. Indeed, let $\{\mathsf{X}_t\}$ be a Gaussian source with power spectrum $S_{\mathsf{X}}(f)$, then the entropy power is defined by

$$P_e(\mathsf{X}) \triangleq \exp\left(\int_{-1/2}^{1/2} \log\left(S(f)\right) \mathrm{d}f\right). \tag{28}$$

According to Wiener's spectral-factorization theory, the entropy power is the MMSE of one-step prediction of X_n from its infinite past, i.e.,

$$P_e(\mathsf{X}) = \inf_{\{b_i\}} \mathbb{E}\left[\left(\mathsf{X}_n - \sum_{i=1}^{\infty} b_i \mathsf{X}_{n-i}\right)\right]. \tag{29}$$

The difference here is that we consider noisy prediction in (27).

The compression scheme in Fig. 3 is based on a corresponding scheme for standard Gaussian rate-distortion problem with MMSE distortion, developed in [14]. The main difference is that the problem considered here is essentially a remote source coding problem with log-loss. In particular, this affects the realization of the shaping filter and the derivation of mutual information fidelity. Our result is as follows:

Theorem 2: Let a bivariate Gaussian stationary source $(\{\mathsf{X}_t, \mathsf{Y}_t\})$ with SNR spectrum $\Gamma(f)$ be given, and assume that $g(\cdot)$ achieves the optimum infinite order prediction error σ_∞. If the quantizer in Fig. 3 is replaced by an AWGN channel then the system of Fig. 3, satisfies

$$\mathcal{R}(C) = \frac{1}{2}\int_{-1/2}^{1/2} \log\left[\frac{1+\Gamma(f)}{1+D_\theta(f)}\right] \mathrm{d}f, \tag{30}$$

where $D_\theta(f)$ is as defined in (15), and θ is chosen such that

$$C = \frac{1}{2}\int_{-1/2}^{1/2} \log\left[\frac{\Gamma(f)}{D_\theta(f)}\right] \mathrm{d}f. \tag{31}$$

Thm. 2 assumes that the quantizer operation is replaced by Gaussian noise. The discrepancy between the ideal Gaussian noise setting and the practical quantizer setting, can be resolved using Vector Quantizer (VQ) and interleaving. Utilizing high-dimensional lattice code for quantization results in nearly Gaussian quantization noise [24]. However, a direct application of a VQ is not well-suited to the sequential prediction scheme of Fig. 3. This problem can be resolved by adding a spatial dimension to the problem, motivated by interleaving for ISI channels [25]. Consider the scheme illustrated in Fig. 4. The input sequence to the encoder is rearranged in matrix form. The columns, whose elements are assumed to be memoryless, are used as the inputs to VQ. The rows on the other hand are fed to the prediction loop. Additional discussion on this technique can be found in [14].

Remark 3: While the central block of the system is sequential and hence *causal*, the whitening filter, shaping filter and the pre- and post filters are noncausal and therefore their realization requires delay. In particular, since $h_{2,n} = h_{1,-n}$, if one of the filters is causal then the other must be anticausal. The problem is that usually the filter's response is infinite,

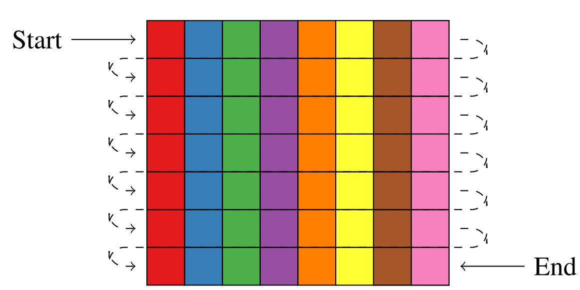

Fig. 4: Source Interleaving Scheme

therefore, the required delay is also unbounded. Indeed, the desired spectrum of the filters may be approximated to any order using filters of sufficiently large but finite delay δ. Then, Thm. 2 holds in general in the limit of infinite delay.

IV. Compound Information Bottleneck

Following [15] we next consider the *compound information bottleneck* (COMIB) function of a bivariate stationary source with memory. Here, the COMIB is given as a limit of normalized mutual information associated with vectors of source samples. For a real valued bivariate source $(\{\mathsf{X}_t, \mathsf{Y}_t\})$ (see (1)), normalized PF constraint C_1, and normalized IB constraint C_2, the COMIB rate can be written as

$$\mathcal{R}^{comib}(C_1, C_2) = \lim_{n\to\infty} R_n^{comib}(C_1, C_2), \tag{32}$$

where

$$R_n^{comib}(C_1, C_2) = \sup_{\mathsf{P}_{\mathsf{Z}^n|\mathsf{Y}^n}} \inf_{\mathsf{P}_{\mathsf{X}^n\mathsf{Y}^n}} \frac{1}{n} I(\mathsf{X}^n; \mathsf{Z}^n) \tag{33}$$

is the corresponding finite vector Gaussian COMIB function. The optimization in (33) is over the sets $\{\mathsf{P}_{\mathsf{X}^n\mathsf{Y}^n}\}$ and $\{\mathsf{P}_{\mathsf{Z}^n|\mathsf{Y}^n}\}$, satisfying $\frac{1}{n}I(\mathsf{X}^n; \mathsf{Y}^n) \geq C_1$ and $\frac{1}{n}I(\mathsf{Y}^n; \mathsf{Z}^n) \leq C_2$, respectively.

A simple way to obtain solution to (32) is by establishing a saddle point property. We briefly remind the reader this property as it will be used in the proof.

Lemma 2 (Optimality of Saddle Point [26, Sec. 5.4.2]): Suppose there exists a saddle point $(\tilde{w}, \tilde{z})$, satisfying $f(\tilde{w}, \tilde{z}) = \inf_{w\in\mathcal{W}} f(w, \tilde{z})$ and $f(\tilde{w}, \tilde{z}) = \sup_{z\in\mathcal{Z}} f(\tilde{w}, z)$, then

$$f(\tilde{w}, \tilde{z}) = \sup_{z\in\mathcal{Z}} \inf_{w\in\mathcal{W}} f(w, z). \tag{34}$$

Next, we consider the respective dual PF problem (7) for Gaussian random variables. The problem in (6) is rather delicate – e.g., if (Y, Z) are scalar jointly Gaussian random variables, the PF rate is zero since one can use the channel from Y to X to describe the less significant bits of Y [27]. Thus, additional constraints should be imposed here to have a non-trivial rate, as stated in the following theorem.

Lemma 3: Suppose $\mathbf{X} \to \mathbf{Y} \to \mathbf{Z}$ constitute a jointly Gaussian vector Markov chain with positive definite marginal covariance matrices $\Sigma_{\mathbf{X}}$, $\Sigma_{\mathbf{Y}}$, and $\Sigma_{\mathbf{Z}}$ respectively, and the cross-covariance matrix of $\mathbf{Z}$ and $\mathbf{Y}$ is given by $\Sigma_{\mathbf{ZY}}$. Let $\Sigma_{\mathbf{YX}}$ be the cross-covariance matrix of the optimal solution to (7). Further, let $U_1\Phi V_1^T$ be the Singular Value Decomposition (SVD) of $\Sigma_{\mathbf{Y}}^{-1/2}\Sigma_{\mathbf{YX}}\Sigma_{\mathbf{X}}^{-1/2}$ and $U_2\Psi V_2^T$ be the SVD of $\Sigma_{\mathbf{Z}}^{-1/2}\Sigma_{\mathbf{ZY}}\Sigma_{\mathbf{Y}}^{-1/2}$. Then, the underlying Gaussian PF problem (7) can be relaxed to the following optimization problem:

$$\begin{aligned} R_n^{\mathsf{PF}}(C_1) = &\min_{U_1\in\mathcal{U}(n), \{\phi_i\}} -\frac{1}{2n} \log\det(I - U_1^T\Phi^2 U_1 V_2^T\Psi^2 V_2) \\ &\text{subject to} \quad -\frac{1}{2n}\sum_{i=1}^{n} \log(1-\phi_i^2) = C_1, \end{aligned} \tag{35}$$

where $\mathcal{U}(n)$ is the set of all $n\times n$ unitary matrices (the unitary group), and $\{\phi_i\}$ are the entries of the diagonal matrix Φ.

Note that in contrast to Lemma 1, we do not have a closed form-solution for the general Gaussian PF problem, and additional numerical optimization is needed. However, for the compound setting we obtain an exact solution which incorporates a white SNR spectrum. Such spectrum has a constant intensity over the entire bandwidth of the signal.

Theorem 3: Suppose that (32) is evaluated for the scenario where $\{\mathsf{X}_t\}$ and $\{\mathsf{Y}_t\}$ are jointly Gaussian random processes with marginal power spectral densities $S_{\mathsf{X}}(f)$ and $S_{\mathsf{Y}}(f)$ respectively. The resulting optimal channel from X_n to Z_n has a linear form, i.e.,

$$\mathsf{Y}_n = h_n * \mathsf{X}_n + \mathsf{W}_n, \tag{36}$$

$$\mathsf{Z}_n = g_n * \mathsf{Y}_n + \mathsf{V}_n, \tag{37}$$

where: h_n and g_n are impulse responses of liner time-invariant filters, whose absolute squared value frequency responses are given by:

$$|H(f)|^2 = \frac{1}{1+\gamma}\frac{S_{\mathsf{Y}}(f)}{S_{\mathsf{X}}(f)}, \quad |G(f)|^2 = \frac{1}{1+\lambda}, \tag{38}$$

$$S_{\mathsf{W}}(f) = \frac{1}{1+\gamma}\frac{S_{\mathsf{Y}}(f)}{S_{\mathsf{X}}(f)}, \quad S_{\mathsf{V}}(f) = \frac{1}{1+\lambda}; \tag{39}$$

$\gamma = 2^{2C_1}-1$; and $\lambda = 2^{2C_2}-1$. Therefore, the optimal double-sided SNR spectrum is white. The corresponding oblivious rate is

$$\mathcal{R}^{comib}(C_1, C_2) = -\frac{1}{2}\log[1-(1-2^{-2C_1})(1-2^{-2C_2})]. \tag{40}$$

This theorem is obtained by applying a saddle-point property from Lemma 2 on Lemma 1 and Lemma 3 and has the following practical implication: The most robust approach in case there is no information regarding the structure of the observed signal and noise is to assume the input is white.

V. Concluding Remarks

In this paper, we have addressed the jointly Gaussian process IB problem. A water-filling type solution has been obtained. Then, a linear prediction scheme that attains the IB was proposed and analyzed. Finally, a closed-form solution has been given to a compound version of the IB for Gaussian processes.

Future research directly related to the results of this paper calls for further investigating single-letter quantization algorithms with information-theoretic metrics. In addition, it would be interesting to consider the IB and PF problems, when the constraint is not $I(\mathsf{Y}; \mathsf{Z}) \leq C$, but $H(\mathsf{Z}) \leq C$, as also done in [20]. Then the transformation $\mathsf{Y} \to \mathsf{Z}$ is known to be discrete.

Acknowledgment

The work has been supported by the European Union's Horizon 2020 Research And Innovation Programme, grant agreement no. 694630, by the ISF under Grant 1791/17, and by the WIN consortium via the Israel minister of economy and science.

References

[1] N. Tishby, F. C. N. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Annu. Allerton Conf. Commun. Control Comput.*, Sep. 1999, p. 368–377.

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

[3] R. Dobrushin and B. Tsybakov, "Information transmission with additional noise," *IRE Trans. Inf. Theory*, vol. 8, no. 5, pp. 293–304, Sep. 1962.

[4] J. Wolf and J. Ziv, "Transmission of noisy information to a noisy receiver with minimum distortion," *IEEE Trans. Inf. Theory*, vol. 16, pp. 406–411, Jul. 1970.

[5] A. Zaidi, I. E. Aguerri, and S. S. (Shitz), "On the information bottleneck problems: Models, connections, applications and information theoretic views," *Entropy*, vol. 22, no. 2, p. 151, Feb. 2020.

[6] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, "Information bottleneck for Gaussian variables," *J. Mach. Learn. Res.*, vol. 6, pp. 165–188, Dec. 2005.

[7] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. 18, no. 4, pp. 460–473, Jul. 1972.

[8] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. 18, pp. 14–20, Jan. 1972.

[9] T. Berger, *Rate distortion theory : A mathematical basis for data compression.*, Englewood Cliffs, N.J, 1971.

[10] W. Hirt and J. Massey, "Capacity of the discrete-time Gaussian channel with intersymbol interference," *IEEE Trans. Inf. Theory*, vol. 34, no. 3, pp. 38–38, May 1988.

[11] S. Shamai, L. Ozarow, and A. Wyner, "Information rates for a discrete-time Gaussian channel with intersymbol interference and stationary inputs," *IEEE Trans. Inf. Theory*, vol. 37, no. 6, pp. 1527–1539, Nov. 1991.

[12] R. M. Gray, "Toeplitz and circulant matrices: A review," *Foundations and Trends® in Communications and Information Theory*, vol. 2, no. 3, pp. 155–239, 2006.

[13] J. Cioffi, G. Dudevoir, M. Vedat Eyuboglu, and G. Forney, "MMSE decision-feedback equalizers and coding. I. Equalization results," *IEEE Trans. Commun.*, vol. 43, no. 10, pp. 2582–2594, Oct. 1995.

[14] R. Zamir, Y. Kochman, and U. Erez, "Achieving the Gaussian rate–distortion function by prediction," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 3354–3364, Jul. 2008.

[15] M. Dikshtein, N. Weinberger, and S. Shamai (Shitz), "The compound information bottleneck outlook," *CoRR*, vol. abs/2205.04567v1, 2022. [Online]. Available: https://arxiv.org/abs/2205.04567v1

[16] ——, "On information bottleneck for Gaussian processes," *CoRR*, vol. abs/2208.10136, 2022. [Online]. Available: https://arxiv.org/abs/2208.10136

[17] A. Globerson and N. Tishby, "On the optimality of the Gaussian information bottleneck curve," *The Hebrew University of Jerusalem, Tech. Rep*, p. 22, 2004.

[18] A. Winkelbauer, S. Farthofer, and G. Matz, "The rate-information trade-off for Gaussian vector channels," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2014, pp. 2849–2853.

[19] M. Meidlinger, A. Winkelbauer, and G. Matz, "On the relation between the Gaussian information bottleneck and MSE-optimal rate-distortion quantization," in *Proc. IEEE SSP*, Jun. 2014, pp. 89–92.

[20] A. Homri, M. Peleg, and S. Shamai Shitz, "Oblivious fronthaul-constrained relay for a Gaussian channel," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5112–5123, Nov. 2018.

[21] A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Médard, "From the information bottleneck to the privacy funnel," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Nov. 2014, pp. 501–505.

[22] 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.

[23] Gibson, *Digital compression for multimedia : principles and standards / Jerry D. Gibson ... [et al.]*. San Francisco, Calif: Morgan Kaufmann, 1998.

[24] R. Zamir, B. Nazer, Y. Kochman, and I. Bistritz, *Lattice Coding for Signals and Networks: A Structured Coding Approach to Quantization, Modulation and Multiuser Information Theory*. Cambridge University Press, 2014.

[25] T. Guess and M. Varanasi, "An information-theoretic framework for deriving canonical decision-feedback receivers in gaussian channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 1, pp. 173–187, Jan. 2005.

[26] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA.: Cambridge University Press, 2014.

[27] S. Shamai, "The information bottleneck: A unified information theoretic view," National Conference on Communications (NCC2021), Jul. 2021, Plenary Address.