Intuitive dissection of the Gaussian information bottleneck method with an application to optimal prediction

Vahe Galstyan,* Age Tjalma, and Pieter Rein ten Wolde†
AMOLF, Science Park 104, 1098 XG Amsterdam, The Netherlands

(Received 22 July 2025; accepted 4 November 2025; published 5 December 2025)

Efficient signal representation is essential for the functioning of living and artificial systems operating under resource constraints. A widely recognized framework for deriving such representations is the information bottleneck method, which yields the optimal strategy for encoding a random variable, such as the signal, in a way that preserves maximal information about a functionally relevant variable, subject to an explicit constraint on the amount of information encoded. While in its general formulation the information bottleneck method is a numerical scheme, it admits an analytical solution in an important special case where the variables involved are jointly Gaussian. In this setting, the solution predicts discrete transitions in the dimensionality of the optimal representation as the encoding capacity is increased. Although these signature transitions, along with other features of the optimal strategy, can be derived from a constrained optimization problem, a clear and intuitive understanding of their emergence is still lacking. In our work, we advance our understanding of the Gaussian information bottleneck method through multiple mutually enriching perspectives, including geometric and information-theoretic ones. These perspectives offer intuition about the set of optimal encoding directions, the nature of the critical points where the optimal number of encoding components changes, and the way the optimal strategy navigates between these critical points. We then apply our treatment of the information bottleneck to a previously studied signal prediction problem, obtaining insights into how different features of the signal are encoded across multiple components to enable optimal prediction of future signals. Altogether, our work deepens the foundational understanding of the information bottleneck method in the Gaussian setting, motivating the exploration of analogous perspectives in broader, non-Gaussian contexts.

DOI: 10.1103/fnd1-8hty

I. INTRODUCTION

Many engineered and natural systems process stimuli received from their environment and subsequently adjust their behavior in response [1–5]. A key step during processing is the creation of an internal representation of the environment via stimulus encoding. Intuitively, more accurate representations are harder to obtain due to stricter requirements on available resources and, potentially, on the architecture of the signal processing mechanism. Since the representation only partially captures the full signal due to limited encoding capacity, the system may choose which of the signal features to store. What features are chosen ultimately depends on their relevance to the tasks of the system.

The information bottleneck method is a general information-theoretic approach for finding optimal representations [6]. In this framework, a stochastic encoding procedure compresses the signal $\mathbf{s}$ into the encoding variable $\mathbf{z}$, both of which are, in general, multidimensional random variables. A key element of the problem that dictates which signal features must be encoded is the so-called relevance variable, denoted here by $\mathbf{y}$. Correlated with the signal, $\mathbf{y}$ is itself a random variable known to have functional importance for the system. In the biological context of cell signaling, for example, the signal $\mathbf{s}$ may represent the time-dependent concentration of nutrients and $\mathbf{z}$ can correspond to the levels of intracellular readout molecules, with the cellular signaling network defining the $\mathbf{s} \to \mathbf{z}$ mapping rule. If the system needs to anticipate the changes in its environment, then $\mathbf{y}$ may represent the future value or future derivative of the stochastic signal, the knowledge of which would enable the system to mount a response in advance [3,7,8]. In the broader context of machine learning, if the goal is to perform signal classification, then $\mathbf{y}$ may represent a low-dimensional signal category label [9]. In all these cases, the optimal encoding strategy should selectively preserve those features of the signal that are maximally informative about the relevance variable.

In the bottleneck method, the cost of encoding the signal is captured via the mutual information $I(\mathbf{z};\mathbf{s})$, while the quality of encoding is captured through $I(\mathbf{z};\mathbf{y})$—a measure of how informative the signal representation $\mathbf{z}$ is about the relevance variable $\mathbf{y}$ [Fig. 1(a)]. The method yields the best encoding strategy via a stochastic $\mathbf{s} \to \mathbf{z}$ mapping rule that maximizes the relevant information $I(\mathbf{z};\mathbf{y})$ for the given amount of

*Contact author: v.galstyan@amolf.nl
†Contact author: tenwolde@amolf.nl

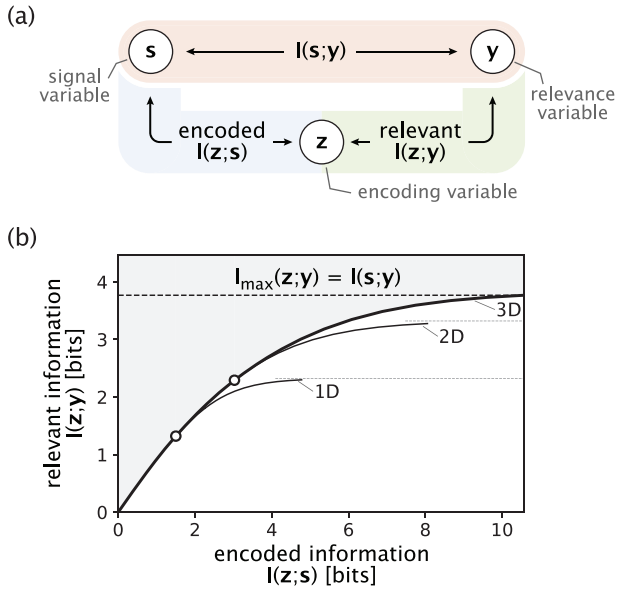

FIG. 1. Elements and operation of the information bottleneck method. (a) Three key variables of the problem, together with the information metrics defined for each pair. $I(\mathbf{z};\mathbf{s})$ is the information encoded in the variable $\mathbf{z}$ about the signal $\mathbf{s}$, and $I(\mathbf{z};\mathbf{y})$ is the information that $\mathbf{z}$ retains about the relevance variable $\mathbf{y}$. Independently of the encoding procedure, intrinsic correlations between $\mathbf{s}$ and $\mathbf{y}$ set the mutual information $I(\mathbf{s};\mathbf{y})$, which serves as an upper bound on the relevant information $I(\mathbf{z};\mathbf{y})$. (b) Information curve arising in the Gaussian information bottleneck method for an example setting with three-dimensional $\mathbf{s}$ and $\mathbf{y}$ variables. The white dots mark the transitions from lower- to higher-dimensional optimal signal representations as the encoding capacity is increased. Thinner curves that saturate at lower levels of relevant information correspond to suboptimal strategies where the dimension of $\mathbf{z}$ is less than 3 and does not vary with increasing encoding capacity. The Lagrange multiplier γ in the objective functional $\mathcal{L} = I(\mathbf{z};\mathbf{y}) - \gamma I(\mathbf{z};\mathbf{s})$ decreases as the curves rise from the origin. In the figure, the three curves terminate at the same value of γ. The gray region above the information curve is inaccessible.

encoded information $I(\mathbf{z};\mathbf{s})$. This problem is typically formulated in the literature as one of constrained optimization where the functional $\mathcal{L}[p(\mathbf{z}|\mathbf{s})] = I(\mathbf{z};\mathbf{y}) - \gamma I(\mathbf{z};\mathbf{s})$ is maximized over encoding strategies $p(\mathbf{z}|\mathbf{s})$, with γ being a Lagrange multiplier. Intuitively, the more the constraint on the encoding capacity is relaxed (smaller γ), the larger will the relevant information become under the optimal strategy. In the $\gamma \to 0$ limit where the encoding capacity is no longer constrained, $\mathbf{z}$ will perfectly capture all signal features correlated with the relevance variable, making the relevant information $I(\mathbf{z};\mathbf{y})$ reach its limit $I(\mathbf{s};\mathbf{y})$ set by the intrinsic correlations between $\mathbf{s}$ and $\mathbf{y}$ variables.

In its general formulation, the information bottleneck method provides the optimal encoding strategy through a numerical iterative algorithm [6]. An analytical procedure, however, becomes available in an important special case, namely, when all random variables of the problem are Gaussian together with their joint distributions [10]. Signals with such statistical properties typically emerge from linear models, which are widely applied in studies of natural systems as well as in engineering [11–14].

In the Gaussian information bottleneck method, the optimal signal representation is achieved via noisy linear encoding. This encoding projects the signal along one or multiple directions, which, in the standard treatment of the problem, correspond to the basis vectors derived in canonical correlation analysis (CCA) [15]. A key feature of the framework is the progressively increasing dimensionality of the optimal representation with the growing capacity to encode the full signal [10]. For example, if the signal and relevance variables are three-dimensional vectors, then the optimal signal representation will be a scalar if the amount of information that can be encoded is less than a certain threshold; once this threshold is crossed, a second encoding component will emerge that will capture an additional linearly independent feature of the signal. As the encoding capacity is increased even further, past another threshold it becomes optimal to incorporate a third encoding component. This suggests that the "information curve" capturing the dependence of the relevant information $I(\mathbf{z};\mathbf{y})$ on the encoded information $I(\mathbf{z};\mathbf{s})$ will be composed of distinct analytic segments, each corresponding to a different dimension of the optimal signal representation [Fig. 1(b)]. While this feature of discrete transitions together with other properties of optimal encoding can be demonstrated mathematically, our understanding of their emergence is still incomplete. For instance, why is it that, instead of using multiple encoding components from the outset, the optimal strategy introduces them one by one at special transition points? And what are the defining properties of these points?

The main aim of the current work is to provide an intuitive understanding of the Gaussian information bottleneck method through a combination of analytical and geometric arguments with clear graphical interpretations. As we will discuss in more detail later, a distinguishing element of our approach that makes many of these intuitive arguments possible is the initial standardization of the marginal distributions $P(\mathbf{s})$ and $P(\mathbf{y})$ that makes the covariance matrices of $\mathbf{s}$ and $\mathbf{y}$ variables diagonal, containing equal entries. Due to this procedure, the structural features of the problem, originally contained in the full joint probability distribution $P(\mathbf{s},\mathbf{y})$, get concentrated in the stochastic $\mathbf{s} \to \mathbf{y}$ and $\mathbf{y} \to \mathbf{s}$ mapping rules set by the conditional distributions $P(\mathbf{y}|\mathbf{s}) = \frac{P(\mathbf{s},\mathbf{y})}{P(\mathbf{s})}$ and $P(\mathbf{s}|\mathbf{y}) = \frac{P(\mathbf{s},\mathbf{y})}{P(\mathbf{y})}$, respectively. While not compromising the generality of treatment, this allows disentangling aspects of the problem that would otherwise be impossible to separate and interpret geometrically.

Additionally, in our treatment we leverage the symmetry in the definition of mutual information to gain deeper insights into optimality from distinct and mutually enriching perspectives. This complements the original approach by Chechik *et al.* [10] where the encoded and relevant information amounts $I(\mathbf{z};\mathbf{s})$ and $I(\mathbf{z};\mathbf{y})$ were computed as reductions in the entropy of the encoding variable $\mathbf{z}$ when the signal $\mathbf{s}$ and the relevance variable $\mathbf{y}$, respectively, were given, i.e., $I(\mathbf{z};\mathbf{s}) = H(\mathbf{z}) - H(\mathbf{z}|\mathbf{s})$ and analogously for $I(\mathbf{z};\mathbf{y})$. Here, we will often consider the alternative "decoding" perspective where these information metrics are interpreted as reductions in the entropy of the respective variable ($\mathbf{s}$ or $\mathbf{y}$) when the

encoding value $\mathbf{z}$ is provided, namely, $I(\mathbf{z};\mathbf{s}) = H(\mathbf{s}) - H(\mathbf{s}|\mathbf{z})$ and similarly for $I(\mathbf{z};\mathbf{y})$.

In our work, we first study the Gaussian information bottleneck method for two-dimensional signal and relevance variables. The complexity of this setting is sufficient for illustrating the key features of optimal encoding related to its directionality and dimensionality. Later, we show how the derived results generalize to higher-dimensional cases.

The remaining of the paper is organized as follows. In Sec. II, we introduce the variables, the linear encoding rule, and the problem of obtaining the optimal signal encoding strategy under constrained information cost. Then in Sec. III, we analyze the one-dimensional (1D) encoding scenario and provide interpretations of optimality from two distinct perspectives (encoding and decoding). We extend our analysis to the case of two-dimensional signal encoding in Sec. IV where we offer intuition on how the information from multiple components gets combined and at what point having two encoding components instead of one becomes the preferred strategy. We end our general discussion of Gaussian information bottleneck in Sec. V by showing how our interpretation applies to a three-dimensional setting and, by extension, to an arbitrary multidimensional case.

To illustrate our outlook on a concrete example, in Sec. VI we apply it to the problem of signal prediction. Specifically, we consider signals generated via a stochastically driven mass-spring model—a canonical setup used in prior studies on prediction [3,7,8,16,17]. There, the signal is represented by the fluctuating position of the mass. The dynamics is such that the current position (x_0) and its time derivative (v_0) fully specify the statistics of the future signal. The goal of the information bottleneck method is to encode these two signal features, namely, $\mathbf{s} = [x_0, v_0]^{\mathrm{T}}$, in such a way that the representation $\mathbf{z}$ is maximally informative about the future pair of features, $\mathbf{y} = [x_\tau, v_\tau]^{\mathrm{T}}$. We obtain insightful analytical results on the optimal encoding strategy and illustrate how it varies with the forecast interval and the dynamical regime of the signal.

We end by summarizing our results and discussing their broader implications in Sec. VII.

II. PROBLEM FORMULATION

Consider a two-component signal $\mathbf{s} = [s_1, s_2]^{\mathrm{T}}$ that has a Gaussian distribution centered at the origin ($\langle \mathbf{s} \rangle = \mathbf{0}$). Each signal value maps stochastically to a different Gaussian variable $\mathbf{y} = [y_1, y_2]^{\mathrm{T}}$ called the relevance variable. Figure 2(a) shows example distributions of these two variables and of the stochastic mappings between them. There, the different multivariate Gaussian distributions are depicted visually via their 1σ-level elliptical contours (set of all points where the probability equals $e^{-1/2}$ times its maximum value). We will be using this geometric way of depicting distributions throughout the paper.

Since the storage of the full signal may be impossible or impractical, we are interested in its partial representation, specifically one in the form of a noisy linear transformation:

$$\mathbf{z} = \mathbf{W}^{\mathrm{T}}\mathbf{s} + \boldsymbol{\xi}. \tag{1}$$

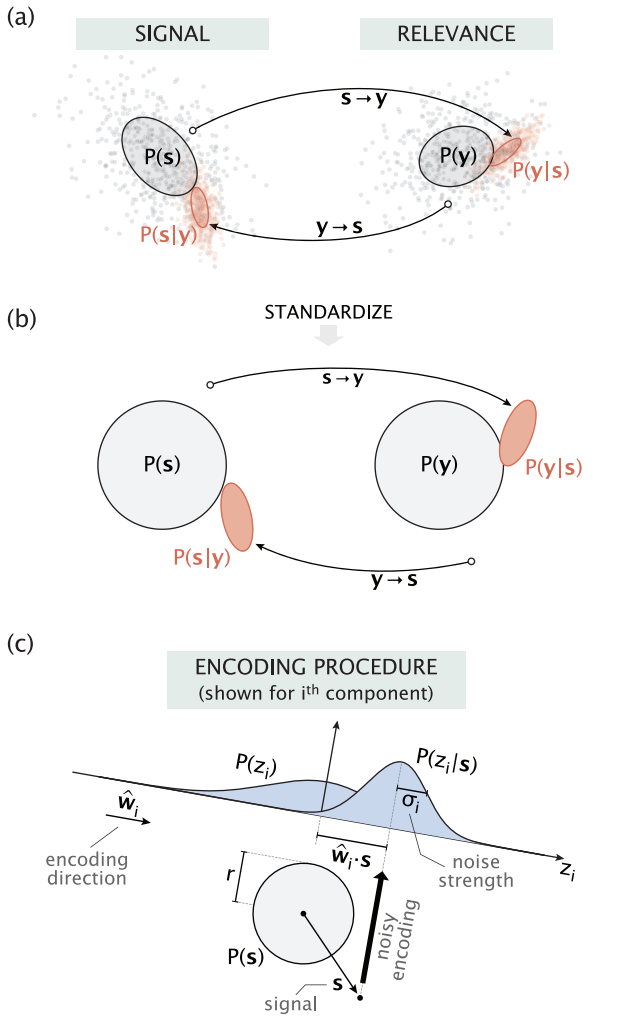

FIG. 2. Visualization of the key probability distributions and procedures in the Gaussian information bottleneck problem. (a) Distributions of the signal and relevance variables, and of the stochastic mappings between them ($\mathbf{s} \to \mathbf{y}$ and $\mathbf{y} \to \mathbf{s}$). The dots are samples from these distributions, while the ellipses represent their 1σ-level contours. (b) Standardization procedure for the distributions $P(\mathbf{s})$ and $P(\mathbf{y})$ that turns their constant-density elliptical contours into circles, thus concentrating key statistical features of the problem in the conditional probabilities $P(\mathbf{s}|\mathbf{y})$ and $P(\mathbf{y}|\mathbf{s})$. (c) Illustration of the encoding procedure for a single component $z_i \sim \mathcal{N}(\hat{\mathbf{w}}_i \cdot \mathbf{s}, \sigma_i^2)$. It can be viewed as a projection of the signal $\mathbf{s}$ onto the encoding direction $\hat{\mathbf{w}}_i$, followed by the addition of the encoding noise ξ_i. Centered at the origin, the marginal distribution $P(z_i)$ has a larger variance $\sigma_{z_i}^2 = r^2 + \sigma_i^2$ that is independent of the encoding direction $\hat{\mathbf{w}}_i$. The cross-correlation matrix used for specifying the statistics of standardized variables in panel (b) is $\boldsymbol{\Sigma}_{\mathbf{sy}} = \begin{pmatrix} 0.83 & -0.50 \\ 0.48 & 0.70 \end{pmatrix}$.

Here, $\mathbf{W}$ is a linear transformation matrix, while $\boldsymbol{\xi}$ is a Gaussian encoding noise vector with zero mean ($\langle \boldsymbol{\xi} \rangle = \mathbf{0}$) and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$. We note that Eq. (1) is, in fact, the optimal form of signal encoding when the signal and relevance variables have jointly Gaussian statistics [10].

As already mentioned in the Introduction, the cost of signal encoding is measured via the mutual information $I(\mathbf{z};\mathbf{s})$. Noisier and hence less accurate representations will have a lower corresponding $I(\mathbf{z};\mathbf{s})$. Conversely, a perfect signal

representation can only be achieved in the limit of infinite encoded information. Multiple encoding strategies defined via the pair $(\mathbf{W}, \mathbf{\Sigma}_{\boldsymbol{\xi}})$ can have the same information cost $I(\mathbf{z};\mathbf{s})$. The goal of Gaussian information bottleneck is to identify the subset of these encoding strategies that remain maximally informative about the relevance variable, i.e., they maximize the relevant information $I(\mathbf{z};\mathbf{y})$ for a given encoding capacity $I(\mathbf{z};\mathbf{s})$ [10].

To help interpret these and other information metrics intuitively, we initially standardize the distributions of signal and relevance variables to make their covariance matrices diagonal, containing equal entries. This procedure does not affect the generality of treatment since the information metrics are invariant under standardization, and the encoding rules for the original variables can be worked out easily through a back transformation (Appendix A1 in the Supplemental Material [18]). Standardization results in the forms $\mathbf{\Sigma}_{\mathbf{s}} = r^2\,\mathbf{I}_{\mathbf{s}}$ and $\mathbf{\Sigma}_{\mathbf{y}} = r^2\,\mathbf{I}_{\mathbf{y}}$, where r^2 is the single-component variance that sets the scale of these variables (same scale chosen for convenience), while $\mathbf{I}_{\mathbf{s}}$ and $\mathbf{I}_{\mathbf{y}}$ are identity matrices with their dimensions given by $\dim(\mathbf{s})$ and $\dim(\mathbf{y})$, respectively. After the standardization procedure, the stochastic mappings $\mathbf{s}\to\mathbf{y}$ and $\mathbf{y}\to\mathbf{s}$ specified via the conditional probability distributions $P(\mathbf{y}|\mathbf{s})$ and $P(\mathbf{s}|\mathbf{y})$, respectively, fully capture the statistical structure of the problem [Fig. 2(b)]. The standardization procedure also makes the ellipses corresponding to these two distributions geometrically identical, although their orientations do not match in general (Appendix A2 in the Supplemental Material [18]). Notably, because the joint statistics of $\mathbf{s}$ and $\mathbf{y}$ variables is Gaussian, the shapes and orientations of $P(\mathbf{s}|\mathbf{y})$ and $P(\mathbf{y}|\mathbf{s})$ ellipses are independent of what specific $\mathbf{y}$ and $\mathbf{s}$ values are used in conditioning.

Of particular interest is the mutual information $I(\mathbf{s};\mathbf{y})$ between the signal and relevance variables. It sets an upper bound on the relevant information $I(\mathbf{z};\mathbf{y})$ since the encoding $\mathbf{z}$ is only a partial representation of the signal. This mutual information can be written in two alternative ways as

$$I(\mathbf{s};\mathbf{y}) = H(\mathbf{s}) - H(\mathbf{s}|\mathbf{y}) \tag{2a}$$

or

$$I(\mathbf{s};\mathbf{y}) = H(\mathbf{y}) - H(\mathbf{y}|\mathbf{s}). \tag{2b}$$

Since all variables of the problem are Gaussian, and the entropy of a d-dimensional Gaussian random variable is of the form $H = \frac{1}{2}\log\left((2\pi e)^d|\mathbf{\Sigma}|\right)$, we can write

$$I(\mathbf{s};\mathbf{y}) = \frac{1}{2}\log\left(\frac{|\mathbf{\Sigma}_{\mathbf{s}}|}{|\mathbf{\Sigma}_{\mathbf{s}|\mathbf{y}}|}\right) = \frac{1}{2}\log\left(\frac{|\mathbf{\Sigma}_{\mathbf{y}}|}{|\mathbf{\Sigma}_{\mathbf{y}|\mathbf{s}}|}\right). \tag{3}$$

Now, it is known that the constant-probability elliptical contour of a multivariate Gaussian distribution has an area that scales with the determinant of the covariance matrix as $\propto |\mathbf{\Sigma}|^{1/2}$. This area sets the entropy of the corresponding distribution. The mutual information $I(\mathbf{s};\mathbf{y})$, viewed as a difference of entropies [Eq. (2b)], can thus be interpreted as a measure of how much the localization space of the relevance variable shrinks upon knowing the signal. For instance, $I(\mathbf{s};\mathbf{y}) = 3$ bits would mean that the area of the ellipse corresponding to $P(\mathbf{y}|\mathbf{s})$ is 8 times smaller than that of the circle corresponding to the standardized marginal distribution $P(\mathbf{y})$. Due to the symmetric definition of mutual information, an analogous interpretation can be made with the variables $\mathbf{s}$ and $\mathbf{y}$ interchanged. In the rest of our work, we will often use this geometric picture for interpreting information measures.

Next, to help illustrate the encoding procedure more clearly, we consider a particular form for the pair of encoding parameters $(\mathbf{W}, \mathbf{\Sigma}_{\boldsymbol{\xi}})$. Specifically, we represent the linear transformation matrix as a collection of unit encoding vectors, i.e., $\mathbf{W} = [\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \ldots]$, and the Gaussian noise vector $\boldsymbol{\xi}$ as one with independent components, implying a diagonal form for its covariance matrix, namely, $\mathbf{\Sigma}_{\boldsymbol{\xi}}^{ij} = \sigma_i^2\delta_{ij}$. We justify this set of considerations in Appendix A3 in the Supplemental Material [18], showing why they do not reduce the generality of the problem treatment. We note that in their original work on the Gaussian information bottleneck, Chechik *et al.* [10] also diagonalized the noise covariance matrix for mathematical convenience. The main difference, however, is that they set $\mathbf{\Sigma}_{\boldsymbol{\xi}}$ equal to an identity matrix and considered tunable amplitudes for the encoding vectors, whereas in our approach we normalize the encoding vectors and treat the noise strengths $\{\sigma_i^2\}$ as distinct tunable parameters. We find that the latter approach, in which the mean encoding variable $\langle\mathbf{z}|\mathbf{s}\rangle = \mathbf{W}^{\mathrm{T}}\mathbf{s}$ stays the same when tuning the noise strengths, yields a more informative geometric picture of the encoding procedure.

Given the above considerations for the pair $(\mathbf{W}, \mathbf{\Sigma}_{\boldsymbol{\xi}})$, individual components of the encoding variable $\mathbf{z}$ can be written as

$$z_i = \hat{\mathbf{w}}_i \cdot \mathbf{s} + \xi_i. \tag{4}$$

Formation of the ith encoding component can thus be interpreted as the projection of the signal vector $\mathbf{s}$ onto the encoding direction $\hat{\mathbf{w}}_i$, followed by the addition of encoding noise ξ_i [Fig. 2(c)]. The conditional distribution of the encoding component z_i is therefore Gaussian with mean $\hat{\mathbf{w}}_i \cdot \mathbf{s}$ and variance σ_i^2, i.e., $P(z_i|\mathbf{s}) \sim \mathcal{N}(\hat{\mathbf{w}}_i \cdot \mathbf{s}, \sigma_i^2)$. Importantly, because we have standardized the signal distribution ($\mathbf{\Sigma}_{\mathbf{s}} = r^2\,\mathbf{I}_{\mathbf{s}}$), the variance of z_i is independent of the corresponding encoding direction $\hat{\mathbf{w}}_i$ and is conveniently given by $\sigma_{z_i}^2 = r^2 + \sigma_i^2$. While this implies that information $I(z_i;\mathbf{s})$ encoded in each component z_i about the signal is also independent of the direction $\hat{\mathbf{w}}_i$, information $I(z_i;\mathbf{y})$ retained in z_i about the relevance variable $\mathbf{y}$ will, in general, depend on $\hat{\mathbf{w}}_i$ and dictate its optimal choice.

The problem of finding the optimal $(\mathbf{W}, \mathbf{\Sigma}_{\boldsymbol{\xi}})$ pair is therefore reduced to obtaining the optimal sets of encoding directions $\{\hat{\mathbf{w}}_i\}$ and corresponding encoding noise strengths $\{\sigma_i^2\}$ that maximize $I(\mathbf{z};\mathbf{y})$ for a given $I(\mathbf{z};\mathbf{s})$. To build intuition on how the optimal strategies emerge, we will first thoroughly study the case of one-dimensional encoding and afterward consider the more general scenarios.

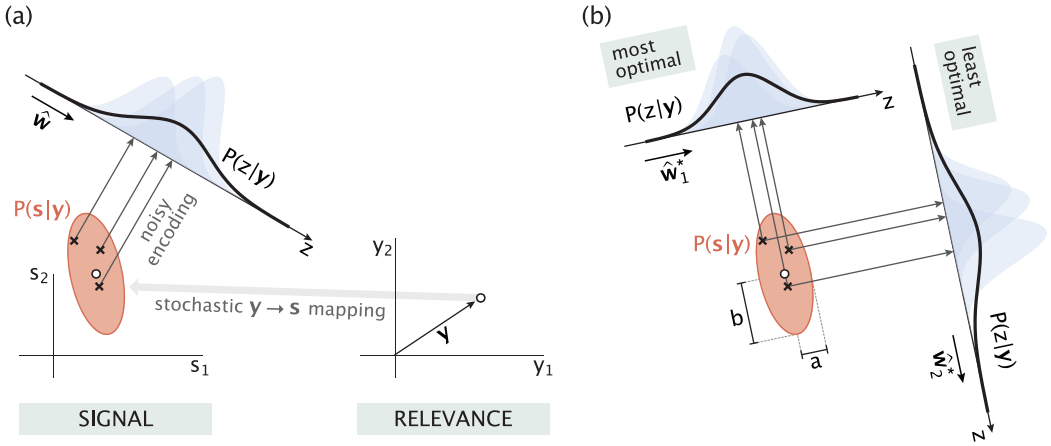

FIG. 3. Graphical perspective on optimality in the one-dimensional encoding setting. (a) Formation of the conditional distribution $P(z|\mathbf{y})$ presented as the projection of $\mathbf{y} \to \mathbf{s}$ mapping statistics $P(\mathbf{s}|\mathbf{y})$ onto the encoding direction $\hat{\mathbf{w}}$, followed by the addition of noise $\xi \sim \mathcal{N}(0, \sigma^2)$. The three example Gaussian distributions shown in blue represent the noisy encoding of different signals $\mathbf{s}$, with the heights of the Gaussians proportional to the corresponding probabilities $P(\mathbf{s}|\mathbf{y})$. (b) Most and least optimal encoding strategies leading to the narrowest and widest distributions $P(z|\mathbf{y})$, respectively.

III. ONE-DIMENSIONAL ENCODING

Suppose the two-component signal $\mathbf{s}$ is encoded in the scalar representation z via

$$z = \hat{\mathbf{w}} \cdot \mathbf{s} + \xi. \tag{5}$$

The mutual information between z and $\mathbf{s}$, which serves as a measure of representation cost, can be computed from the definition $I(z;\mathbf{s}) = H(z) - H(z|\mathbf{s})$ as

$$I(z;\mathbf{s}) = \frac{1}{2}\log\left(\frac{\sigma_z^2}{\sigma_{z|\mathbf{s}}^2}\right) = \frac{1}{2}\log\left(1 + \frac{r^2}{\sigma^2}\right). \tag{6}$$

Here, we substituted the variances of distributions $P(z)$ and $P(z|\mathbf{s})$ plotted in Fig. 2(c). As can be seen, $I(z;\mathbf{s})$ depends on the encoding noise but not on the encoding direction. This means that fixing $I(z;\mathbf{s})$ will fix σ^2, and hence, maximization of the relevant information $I(z;\mathbf{y})$ under fixed $I(z;\mathbf{s})$ must be done by optimally choosing the encoding vector $\hat{\mathbf{w}}$.

Using the definition $I(z;\mathbf{y}) = H(z) - H(z|\mathbf{y})$, we write the relevant information as

$$I(z;\mathbf{y}) = \frac{1}{2}\log\left(\frac{\sigma_z^2}{\sigma_{z|\mathbf{y}}^2}\right). \tag{7}$$

We already know that $\sigma_z^2 = r^2 + \sigma^2$, which is constant for a given $I(z;\mathbf{s})$. To understand what factors contribute to the conditional variance $\sigma_{z|\mathbf{y}}^2$, we represent the corresponding distribution $P(z|\mathbf{y})$ as

$$P(z|\mathbf{y}) = \int d\mathbf{s}\, \underbrace{P(z|\mathbf{s})}_{\substack{\text{noisy}\\ \text{encoding}}}\, \underbrace{P(\mathbf{s}|\mathbf{y})}_{\substack{\text{stochastic}\\ \mathbf{y}\to\mathbf{s}\\ \text{mapping}}}. \tag{8}$$

The two conceptually distinct probabilities inside the integral were already discussed in the previous section; specifically, $P(z|\mathbf{s})$ stands for the noisy encoding procedure for a given signal $\mathbf{s}$ [Fig. 2(c)], while $P(\mathbf{s}|\mathbf{y})$ captures the stochastic $\mathbf{y} \to \mathbf{s}$ mapping [Fig. 2(b)] that does not depend on $\hat{\mathbf{w}}$. The formation of $P(z|\mathbf{y})$ can thus be interpreted mechanistically as the projection of the distribution $P(\mathbf{s}|\mathbf{y})$ onto the encoding direction $\hat{\mathbf{w}}$, followed by the addition of an independent encoding noise [Fig. 3(a)]. This translates into the following expression for the conditional variance:

$$\sigma_{z|\mathbf{y}}^2 = \hat{\mathbf{w}}^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathbf{s}|\mathbf{y}} \hat{\mathbf{w}} + \sigma^2. \tag{9}$$

The first term on the right-hand side depends on the encoding direction $\hat{\mathbf{w}}$ and therefore can be tuned, while the second term (σ^2) is fixed for given $I(z;\mathbf{s})$.

Now, we know from Eq. (7) that the relevant information is the largest when $\sigma_{z|\mathbf{y}}^2$ is minimal or, equivalently, the distribution $P(z|\mathbf{y})$ is the narrowest. In view of Fig. 3(a), the condition for minimizing the width of the $P(z|\mathbf{y})$ distribution becomes straightforward: the encoding vector $\hat{\mathbf{w}}$ must be parallel to the minor axis of the $\mathbf{y} \to \mathbf{s}$ mapping ellipse. This means that the optimal $\hat{\mathbf{w}}$ is an eigenvector of the covariance matrix $\boldsymbol{\Sigma}_{\mathbf{s}|\mathbf{y}}$ with the smaller corresponding eigenvalue, i.e., $\boldsymbol{\Sigma}_{\mathbf{s}|\mathbf{y}} \hat{\mathbf{w}}_1^* = a^2 \hat{\mathbf{w}}_1^*$, where a is the length of the semiminor axis (see Appendix B1 in the Supplemental Material [18] for details). Similarly, the least favorable encoding direction $\hat{\mathbf{w}}_2^*$ is parallel to the major axis of the ellipse, satisfying the criterion $\boldsymbol{\Sigma}_{\mathbf{s}|\mathbf{y}} \hat{\mathbf{w}}_2^* = b^2 \hat{\mathbf{w}}_2^*$, where b is the semimajor axis length ($b \geqslant a$). These two options are shown in Fig. 3(b).

Using the optimal encoding direction $\hat{\mathbf{w}} = \hat{\mathbf{w}}_1^*$ in Eq. (9) and substituting the resulting expression for $\sigma_{z|\mathbf{y}}^2$ into Eq. (7), we find the relevant information under optimal encoding to be

$$I^{\mathrm{opt}}(z;\mathbf{y}) = \frac{1}{2}\log\left(\frac{r^2 + \sigma^2}{a^2 + \sigma^2}\right). \tag{10}$$

As expected, $I^{\mathrm{opt}}(z;\mathbf{y})$ is close to zero when the encoding noise dominates the signal ($\sigma \gg r$). Conversely, it is the largest in the limit of noiseless encoding ($\sigma \to 0$) and is given by $I_{\max}^{\mathrm{opt}}(z;\mathbf{y}) = -\log \tilde{a}$, where $\tilde{a} = a/r$ is the normalized semiminor axis length ($0 < \tilde{a} < 1$).

Figure 4 shows the plot of the relevant information $I^{\mathrm{opt}}(z;\mathbf{y})$ under the optimal strategy as a function of the

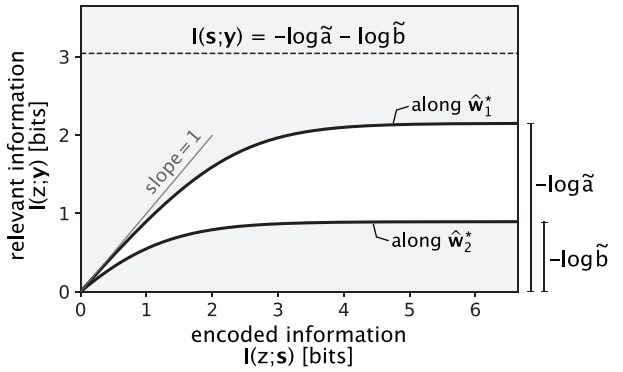

FIG. 4. Relevant vs encoded information in the scalar encoding setting. The two curves correspond to the most and least optimal strategies with encoding directions $\hat{\mathbf{w}} = \hat{\mathbf{w}}_1^*$ and $\hat{\mathbf{w}} = \hat{\mathbf{w}}_2^*$, respectively. Inaccessible regions of the information plane are colored in gray.

encoded information $I(z;\mathbf{s})$. Since the three variables of the bottleneck problem form a Markov chain ($\mathbf{z} - \mathbf{s} - \mathbf{y}$, implying that $\mathbf{z}$ is conditionally independent of $\mathbf{y}$ when given $\mathbf{s}$), the data processing inequality constrains the initial slope of the information curve to be less than 1 [10]. In fact, this slope, given by $1 - \tilde{a}^2$ (see Appendix B2 in the Supplemental Material [18]), reflects the strong data processing inequality at the onset of learning where the increase in the relevant information per encoded bit is the largest $[\partial I(z;\mathbf{y})/\partial I(z;\mathbf{s})|_{I(z;\mathbf{s})=0} = \gamma_c = 1 - \tilde{a}^2$, where γ_c is the critical value of the Lagrange multiplier representing the initial slope] [19]. Below the optimal curve, we also plot $I(z;\mathbf{y})$ vs $I(z;\mathbf{s})$ for the least favorable encoding strategy with $\hat{\mathbf{w}} = \hat{\mathbf{w}}_2^*$. There, the curve has an initial slope of $1 - \tilde{b}^2$ and saturates at the value $-\log \tilde{b}$.

Since both $\mathbf{s}$ and $\mathbf{y}$ are multidimensional, the scalar encoding variable z cannot capture all the information that the full signal $\mathbf{s}$ contains about relevance variable $\mathbf{y}$, even in the noiseless encoding limit. This information, expressed in terms of the parameters $\tilde{a}$ and $\tilde{b}$, can be written as $I(\mathbf{s};\mathbf{y}) = -\log \tilde{a} - \log \tilde{b}$. It is shown in Fig. 4 as an unattainable bound. As the figure hints, this bound can be reached when a second encoding component with maximum information contribution $-\log \tilde{b}$ is combined with the first component with contribution $-\log \tilde{a}$. In Sec. IV, we will study this scenario in detail.

Alternative "decoding" perspective

Before considering the two-dimensional encoding scenario, we present here an alternative perspective on optimality that will complement our understanding in the scalar encoding case and later serve as a key framework for studying the higher-dimensional cases. We start off by revisiting the encoded information, this time based on the definition $I(z;\mathbf{s}) = H(\mathbf{s}) - H(\mathbf{s}|z)$, rather than $I(z;\mathbf{s}) = H(z) - H(z|\mathbf{s})$ used in the previous section. It represents the reduction in the uncertainty about the signal achieved when knowing the encoding value. Using the encoding rule in Eq. (5), we can derive the statistical properties of the "decoding" distribution $P(\mathbf{s}|z)$ (see Appendix B3 in the Supplemental Material [18]). Specifically, its mean is given by

$$\langle \mathbf{s}|z \rangle = \frac{r^2}{r^2 + \sigma^2} \hat{\mathbf{w}} z, \tag{11}$$

while the conditional covariance matrix is

$$\mathbf{\Sigma}_{\mathbf{s}|z} = r^2 \left(\mathbf{I}_\mathbf{s} - \frac{r^2}{r^2 + \sigma^2} \hat{\mathbf{w}} \hat{\mathbf{w}}^\mathrm{T} \right). \tag{12}$$

As can be seen from Eq. (11) and illustrated in Fig. 5(a), the mean vector $\langle \mathbf{s}|z \rangle$ is parallel to the encoding direction $\hat{\mathbf{w}}$, with its magnitude dependent both on the encoding value z and on the encoding noise strength σ. In the noiseless limit ($\sigma \to 0$), it has the largest magnitude (equal to $\hat{\mathbf{w}}z$) and ends on the constant-z line (the line perpendicular to $\hat{\mathbf{w}}$ that passes through $\mathbf{s} = \hat{\mathbf{w}}z$), while in the infinite noise limit ($\sigma \to \infty$), it reaches the origin, retaining no information about the signal.

Although the position of the ellipse corresponding to $P(\mathbf{s}|z)$ depends on the encoding value z, the shape of the ellipse does not (a property of Gaussian statistics); the latter only depends on the signal magnitude r and the encoding noise level σ. The ellipse is compressed along the encoding direction $\hat{\mathbf{w}}$, with its semiminor axis given by $\ell = r\sigma/\sqrt{r^2 + \sigma^2}$. In the noiseless limit ($\sigma \to 0$), the ellipse gets localized on the constant-z line. Although the corresponding encoded information $I(z;\mathbf{s})$ is infinite in this limit [due to the $P(\mathbf{s}|\mathbf{z})$ ellipse having zero area], recovering the full vector signal $\mathbf{s}$ remains impossible since scalar encoding only informs on a single projection of $\mathbf{s}$, providing no information in the perpendicular direction. In the opposite limit ($\sigma \to \infty$), information about the signal is obscured completely and the ellipse turns into a circle centered at the origin.

We next use a similar approach to interpret the relevant information, writing it as $I(z;\mathbf{y}) = H(\mathbf{y}) - H(\mathbf{y}|z)$. The marginal distribution $P(\mathbf{y})$, as discussed earlier, is standardized and is presented graphically as a circle of radius r [Fig. 5(b)]. To better understand the conditional distribution $P(\mathbf{y}|z)$ corresponding to the second entropy term $H(\mathbf{y}|z)$, we write

$$P(\mathbf{y}|z) = \int \mathrm{d}\mathbf{s} \underbrace{P(\mathbf{s}|z)}_{\substack{\text{noisy} \\ \text{decoding}}} \underbrace{P(\mathbf{y}|\mathbf{s})}_{\substack{\text{stochastic} \\ \mathbf{s} \to \mathbf{y} \\ \text{mapping}}}. \tag{13}$$

In view of the above integral expression, the formation of $P(\mathbf{y}|z)$ can be understood as the stochastic mapping of the decoded signal distribution $P(\mathbf{s}|z)$ onto the relevance plane. Derived in Appendix B3 in the Supplemental Material [18], the mean of this distribution is

$$\langle \mathbf{y}|z \rangle = \frac{r^2}{r^2 + \sigma^2} \mathbf{v} z, \tag{14}$$

and the covariance matrix is given by

$$\mathbf{\Sigma}_{\mathbf{y}|z} = r^2 \left(\mathbf{I}_\mathbf{y} - \frac{r^2}{r^2 + \sigma^2} \mathbf{v}\mathbf{v}^\mathrm{T} \right). \tag{15}$$

Here, we have introduced the vector

$$\mathbf{v} = \tilde{\mathbf{\Sigma}}_{\mathbf{ys}} \hat{\mathbf{w}}, \tag{16}$$

with $\tilde{\mathbf{\Sigma}}_{\mathbf{ys}} = \mathbf{\Sigma}_{\mathbf{ys}}/r^2$ and $||\mathbf{v}|| \leqslant 1$. It depends both on the encoding direction $\hat{\mathbf{w}}$ and on the $\mathbf{s} \leftrightarrow \mathbf{y}$ statistics characterized

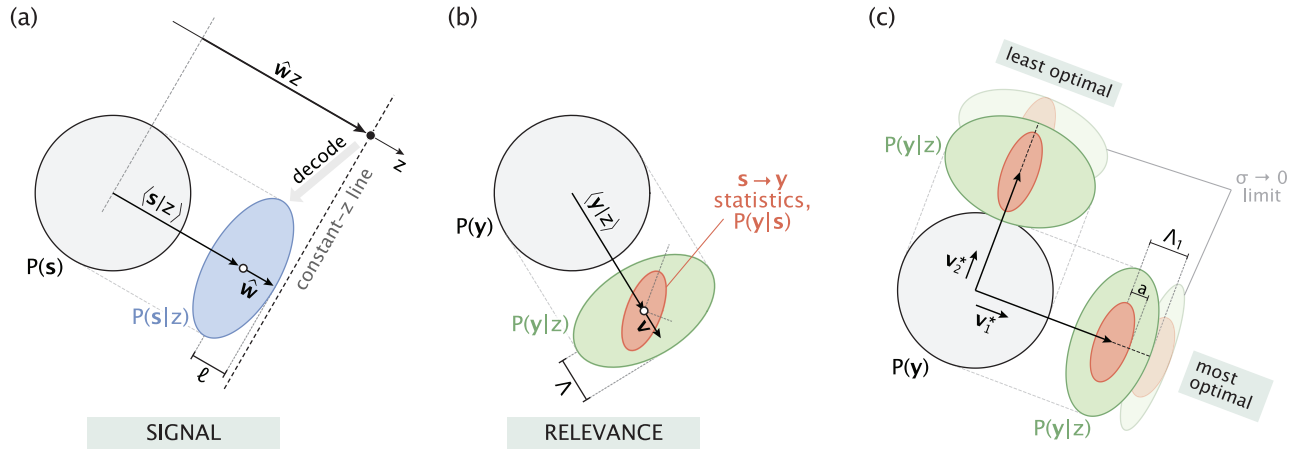

FIG. 5. Alternative perspective on optimality in the scalar encoding scenario. (a) Decoding of the signal $\mathbf{s}$ from the encoding variable z. The decoding distribution $P(\mathbf{s}|z)$ collapses onto the constant-z line in the $\sigma \to 0$ limit. (b) Formation of the distribution $P(\mathbf{y}|z)$ viewed as the stochastic mapping of $P(\mathbf{s}|z)$ onto the relevance plane. The label "$\mathbf{s} \to \mathbf{y}$ statistics" represents the distribution $P(\mathbf{y} \mid \mathbf{s} = \langle \mathbf{s}|z \rangle)$. (c) The distribution $P(\mathbf{y}|z)$ arising under the most and least optimal strategies (shown as green ellipses), overlaid by the ellipses capturing the $\mathbf{s} \to \mathbf{y}$ mapping statistics (shown in red). Transparent ellipses represent the noiseless encoding limit ($\sigma \to 0$). $\mathbf{v}_1^*$ and $\mathbf{v}_2^*$ are the directions on the $\mathbf{y}$ plane along which information about the relevance variable is preserved under the two respective strategies. No information on $\mathbf{y}$ is preserved in directions perpendicular to the $\mathbf{v}$ vectors, which is the reason why the semimajor axes of $P(\mathbf{y}|z)$ ellipses are not compressed and have a length equal to r—the radius of the $P(\mathbf{y})$ circle. In making the plots, the encoding value z was kept the same.

by $\tilde{\mathbf{\Sigma}}_{\mathbf{ys}}$. Functionally, the vector $\mathbf{v}$ sets the direction on the relevance plane along which information about the $\mathbf{y}$ variable is preserved in the encoding value z. This property is reflected in the shape of the ellipse corresponding to the distribution $P(\mathbf{y}|z)$, which is compressed along the direction of $\mathbf{v}$ [Fig. 5(b)]. When decreasing the encoding noise σ, the $P(\mathbf{y}|z)$ ellipse will get more compressed along $\mathbf{v}$. However, in the direction perpendicular to $\mathbf{v}$, the size of the $P(\mathbf{y}|z)$ ellipse will remain unchanged and equal to that of the $P(\mathbf{y})$ circle because z stores no information about $\mathbf{y}$ in that direction.

Now, independent of the encoding strategy, there exists a direction on the $\mathbf{y}$ plane along which the uncertainty about the relevance variable gets reduced the most upon knowing the signal value $\mathbf{s}$. This direction is set by the eigenvector of the covariance matrix $\mathbf{\Sigma}_{\mathbf{y}|\mathbf{s}}$ with the smaller corresponding eigenvalue, which is along the minor axis of the ellipse representing the $\mathbf{s} \to \mathbf{y}$ mapping statistics [Fig. 5(b)].

For an arbitrary choice of $\hat{\mathbf{w}}$, the direction along which relevant information is preserved ($\mathbf{v}$) does not match the direction of least uncertainty in the $\mathbf{s} \to \mathbf{y}$ mapping [Fig. 5(b)]. Under optimal encoding (with $\hat{\mathbf{w}} = \hat{\mathbf{w}}_1^*$), however, these directions have to match (see Appendix B4 in the Supplemental Material [18] for the proof). This optimality feature is illustrated in Fig. 5(c) where the ellipses corresponding to the decoded relevance distribution $P(\mathbf{y}|z)$ and $\mathbf{s} \to \mathbf{y}$ statistics have their minor axes aligned. The semiminor axis of the optimal $P(\mathbf{y}|z)$ ellipse has a length equal to $\Lambda_1 = r\sqrt{(a^2+\sigma^2)/(r^2+\sigma^2)}$, which approaches r as the encoding noise σ becomes very large and converges to a in the limit of noiseless encoding. The result derived earlier for the maximum relevant information under fixed encoded information [Eq. (10)] follows directly from this semiminor axis expression via $I^{\text{opt}}(z;\mathbf{y}) = \log(r/\Lambda_1)$, where r/Λ_1 is the area ratio of ellipses corresponding to the marginal distribution $P(\mathbf{y})$ and the optimal $P(\mathbf{y}|z)$, respectively.

In an analogous way, one can show that the least optimal strategy preserves information about the relevance variable in the most uncertain direction of $\mathbf{s} \to \mathbf{y}$ mapping, which is along the major axis of the mapping ellipse [Fig. 5(c)]. Notably, just as the most and least optimal encoding directions, namely, $\hat{\mathbf{w}}_1^*$ and $\hat{\mathbf{w}}_2^*$ [see Fig. 3(b)], are perpendicular to one another, so are the corresponding vectors $\mathbf{v}_1^*$ and $\mathbf{v}_2^*$ [Fig. 5(c)]. This set of properties will be very useful for extending the perspective developed here to the two-dimensional encoding scenario, which we will study next.

IV. TWO-DIMENSIONAL ENCODING

Earlier in Fig. 4, we saw that a scalar representation z is unable to fully encode the mutual information $I(\mathbf{s};\mathbf{y})$ between two-dimensional signal and relevance variables. In this section, we will consider the option of encoding the signal into a two-dimensional vector, will understand when it becomes preferred over the scalar encoding option, and how it allows reaching the bound on relevant information given by $I(\mathbf{s};\mathbf{y})$.

Vector encoding means that we now have two encoding components defined as $z_1 = \hat{\mathbf{w}}_1 \cdot \mathbf{s} + \xi_1$ and $z_2 = \hat{\mathbf{w}}_2 \cdot \mathbf{s} + \xi_2$. These components are specified via encoding directions $\{\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2\}$ and encoding noise strengths $\{\sigma_1^2, \sigma_2^2\}$. The optimal encoding problem is about finding optimal choices of these four parameters that maximize the relevant information $I(\mathbf{z};\mathbf{y})$ under fixed encoded information $I(\mathbf{z};\mathbf{s})$.

We start off by considering a natural choice for the optimal directions $\hat{\mathbf{w}}_1$ and $\hat{\mathbf{w}}_2$ consisting of the optimal 1D encoding direction $\hat{\mathbf{w}}_1^*$ and the direction perpendicular to it, $\hat{\mathbf{w}}_2^*$, shown in Fig. 3(b). Under this assignment, z_1 and z_2 encode linearly independent projections of the signal and,

furthermore, retain information about the relevance variable along perpendicular directions [$\mathbf{v}_1^*$ and $\mathbf{v}_2^*$; see Fig. 5(c)]. Notably, the directions $\{\hat{\mathbf{w}}_i^*, \hat{\mathbf{v}}_i^*\}$ correspond to the basis vectors used in canonical correlation analysis—a statistical method that identifies maximally correlated pairs of linear projections for two multidimensional variables [15]. In CCA, $\hat{\mathbf{w}}_1^* \cdot \mathbf{s}$ and $\hat{\mathbf{v}}_1^* \cdot \mathbf{y}$ are the most highly correlated projections of the signal and relevance variables, respectively, while $\hat{\mathbf{w}}_2^* \cdot \mathbf{s}$ and $\hat{\mathbf{v}}_2^* \cdot \mathbf{y}$ are the second-most correlated projections that are linearly independent of the first pair. This parallel was drawn also in the original work by Chechik *et al.* [10]. We discuss the degenerate space of optimal solutions briefly at the end of the next section and in greater detail in Appendix D2 in the Supplemental Material [18], where we show how correlated z_i components that encode along nonperpendicular directions can also yield equally optimal solutions. In the rest of this section, we will focus on the problem of optimally assigning the encoding noise strengths $\{\sigma_1^2, \sigma_2^2\}$ to the principal directions $\hat{\mathbf{w}}_1 = \hat{\mathbf{w}}_1^*$ and $\hat{\mathbf{w}}_2 = \hat{\mathbf{w}}_2^*$ under a constraint of fixed $I(\mathbf{z};\mathbf{s})$, which is at the heart of the Gaussian information bottleneck method and is a problem not addressed in canonical correlation analysis.

Following the approach developed in the previous section, we examine the formation of the decoding distribution $P(\mathbf{s}|\mathbf{z})$. The mean of this distribution, given by

$$\langle \mathbf{s}|\mathbf{z}\rangle = \sum_{i\in\{1,2\}} \frac{r^2}{r^2+\sigma_i^2}\,\hat{\mathbf{w}}_i z_i, \tag{17}$$

represents the sum of mean signals decoded from the two separate $\mathbf{z}$ components, i.e., $\langle \mathbf{s}|\mathbf{z}\rangle = \sum_i \langle \mathbf{s}|z_i\rangle$. Note that the scalar encoding case is recovered in the $\sigma_2 \to \infty$ limit, since $\langle \mathbf{s}|z_2\rangle = \mathbf{0}$ in that limit. The covariance matrix of $P(\mathbf{s}|\mathbf{z})$, which is of the form

$$\mathbf{\Sigma}_{\mathbf{s}|\mathbf{z}} = r^2\left(\mathbf{I}_\mathbf{s} - \sum_{i\in\{1,2\}} \frac{r^2}{r^2+\sigma_i^2}\,\hat{\mathbf{w}}_i\hat{\mathbf{w}}_i^{\mathrm{T}}\right), \tag{18}$$

also reduces to the scalar encoding result [Eq. (12)] in the limit $\sigma_2 \to \infty$ (see Appendix C1 in the Supplemental Material [18] for details).

Figure 6(a) illustrates the formation of the decoding distribution $P(\mathbf{s}|\mathbf{z})$. As can be seen, the ellipse corresponding to $P(\mathbf{s}|\mathbf{z})$ is compressed along both of its axes, which have half-lengths $\ell_i = r\sigma_i/\sqrt{r^2+\sigma_i^2}$. These lengths are the semiminor axis lengths of the scalar encoding ellipses corresponding to the distributions $P(\mathbf{s}|z_1)$ and $P(\mathbf{s}|z_2)$. Note that in the limit $\sigma_2 \to \infty$ we have $\ell_2 \to r$, which results in the $P(\mathbf{s}|z_2)$ ellipse becoming a circle centered at the origin and the $P(\mathbf{s}|\mathbf{z})$ ellipse becoming identical to the $P(\mathbf{s}|z_1)$ ellipse.

Formation of the distribution $P(\mathbf{y}|\mathbf{z})$ on the relevance plane takes place in an analogous way [Fig. 6(b)]. The ellipse corresponding to $P(\mathbf{y}|\mathbf{z})$ has its axes oriented along the vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ (with $\mathbf{v}_i = \tilde{\mathbf{\Sigma}}_{\mathbf{ys}}\hat{\mathbf{w}}_i$). Akin to Eqs. (17) and (18), the mean of $P(\mathbf{y}|\mathbf{z})$ is given by

$$\langle \mathbf{y}|\mathbf{z}\rangle = \sum_{i\in\{1,2\}} \frac{r^2}{r^2+\sigma_i^2}\,\mathbf{v}_i z_i \tag{19}$$

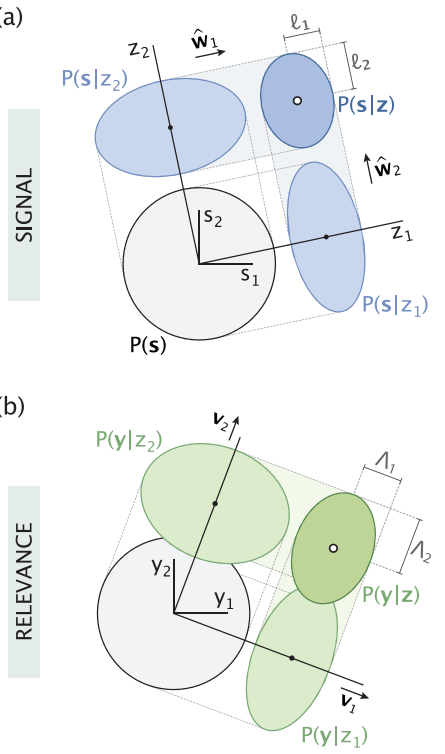

FIG. 6. Formation of the distributions $P(\mathbf{s}|\mathbf{z})$ and $P(\mathbf{y}|\mathbf{z})$ [panels (a) and (b), respectively] in the two-dimensional encoding scenario with optimal choices of perpendicular encoding directions, namely, $\hat{\mathbf{w}}_1 = \hat{\mathbf{w}}_1^*$ and $\hat{\mathbf{w}}_2 = \hat{\mathbf{w}}_2^*$. The corresponding $\mathbf{v}$ vectors (with $\mathbf{v}_i = \tilde{\mathbf{\Sigma}}_{\mathbf{ys}}\hat{\mathbf{w}}_i$) are also perpendicular to each other.

and its covariance matrix is

$$\mathbf{\Sigma}_{\mathbf{y}|\mathbf{z}} = r^2\left(\mathbf{I}_\mathbf{y} - \sum_{i\in\{1,2\}} \frac{r^2}{r^2+\sigma_i^2}\,\mathbf{v}_i\mathbf{v}_i^{\mathrm{T}}\right). \tag{20}$$

The eigenvalues of $\mathbf{\Sigma}_{\mathbf{y}|\mathbf{z}}$ corresponding to the eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are Λ_1^2 and Λ_2^2, respectively, where Λ_1 and Λ_2 represent the half-lengths of the $P(\mathbf{y}|\mathbf{z})$ ellipse. They satisfy the relation $\Lambda_i^2 = r^2(1 - \frac{r^2}{r^2+\sigma_i^2}|\mathbf{v}_i|^2)$. Substituting the identities $|\mathbf{v}_1|^2 = 1-\tilde{a}^2$ and $|\mathbf{v}_2|^2 = 1-\tilde{b}^2$ (see Appendix C1 in the Supplemental Material [18]), one obtains $\Lambda_1 = r\sqrt{(a^2+\sigma_1^2)/(r^2+\sigma_1^2)}$ and $\Lambda_2 = r\sqrt{(b^2+\sigma_2^2)/(r^2+\sigma_2^2)}$. Notably, in the limit of noiseless encoding ($\sigma_1, \sigma_2 \to 0$), Λ_1 and Λ_2 converge to the semiaxis lengths a and b of the $P(\mathbf{y}|\mathbf{s})$ ellipse, respectively.

Optimal noise allocation and dimensionality of representation

In view of Fig. 6(a), fixing the encoded information $I(\mathbf{s};\mathbf{z})$ in the information bottleneck problem means fixing the area of the $P(\mathbf{s}|\mathbf{z})$ ellipse, which is equivalent to fixing $\ell_1 \times \ell_2$—the product of the ellipse's major and minor semiaxis lengths. Crucially, there are many ways of choosing ℓ_1 and ℓ_2 (via σ_1 and σ_2 assignments) that keep the product $\ell_1 \times \ell_2$ and hence $I(\mathbf{s};\mathbf{z})$ the same.

A range of possible options is shown in Fig. 7(a) for an example case with $I(\mathbf{z};\mathbf{s}) = 2.5$ bits of encoded information.

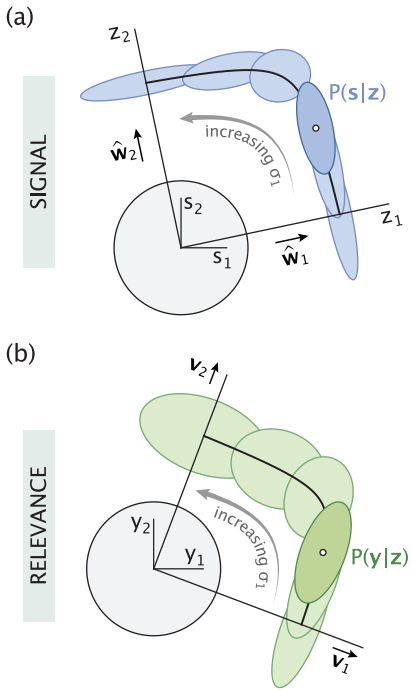

FIG. 7. Distributions $P(\mathbf{s}|\mathbf{z})$ and $P(\mathbf{y}|\mathbf{z})$ [panels (a) and (b), respectively] for different $\{\sigma_1, \sigma_2\}$ assignment options with fixed encoded information $I(\mathbf{z};\mathbf{s}) = 2.5$ bits [increase in σ_1 is accompanied by a simultaneous decrease in σ_2 to keep the area of $P(\mathbf{s}|\mathbf{z})$ ellipses unchanged]. Ellipses in each panel corresponding to the optimal noise assignment [maximizing the relevant information $I(\mathbf{z};\mathbf{y})$ by minimizing the $P(\mathbf{y}|\mathbf{z})$ ellipse area] are drawn in darker colors. Curved lines in both panels represent the mean vectors $\langle \mathbf{s}|\mathbf{z}\rangle$ and $\langle \mathbf{y}|\mathbf{z}\rangle$ drawn when the noise levels are continuously tuned. During noise tuning, the encoding values are kept fixed at $z_1 = \hat{\mathbf{w}}_1 \cdot \mathbf{s}$ and $z_2 = \hat{\mathbf{w}}_2 \cdot \mathbf{s}$ for a specified signal $\mathbf{s}$.

There, the noise assignments vary from finite σ_1 and infinite σ_2 (all information contained in the z_1 component) to infinite σ_1 and finite σ_2 (all information contained in the z_2 component), with different intermediate cases in between (both σ_1 and σ_2 finite). Now, each $\{\sigma_1, \sigma_2\}$ assignment option also has its corresponding $P(\mathbf{y}|\mathbf{z})$ distribution on the relevance plane [Fig. 7(b)]. In contrast to $P(\mathbf{s}|\mathbf{z})$ ellipses that all have the same area [due to encoded information $I(\mathbf{z};\mathbf{s})$ being fixed], the $P(\mathbf{y}|\mathbf{z})$ ellipses generally have different areas, which reflects the fact that the relevant information $I(\mathbf{y};\mathbf{z})$ depends on the noise assignment strategy. The optimal strategy is the one that maximizes $I(\mathbf{z};\mathbf{y})$ by minimizing the area of the corresponding $P(\mathbf{y}|\mathbf{z})$ ellipse for a given amount of $I(\mathbf{s};\mathbf{z})$.

In Fig. 7, distributions $P(\mathbf{s}|\mathbf{z})$ and $P(\mathbf{y}|\mathbf{z})$ corresponding to the optimal noise assignment are shown highlighted. As can be seen, the optimal $P(\mathbf{s}|\mathbf{z})$ ellipse is more compressed along the $\hat{\mathbf{w}}_1$ direction compared to the $\hat{\mathbf{w}}_2$ direction, and similarly, the corresponding $P(\mathbf{y}|\mathbf{z})$ ellipse is more compressed along $\mathbf{v}_1$ compared to $\mathbf{v}_2$. This reflects the intuitive expectation that the optimal strategy would prioritize the more informative $\hat{\mathbf{w}}_1$ direction over the less informative $\hat{\mathbf{w}}_2$ direction through a biased allocation of encoding noises, i.e., $\sigma_1 \leqslant \sigma_2$ (hence, $\ell_1 \leqslant \ell_2$).

By what principle is this bias set? The answer is illustrated in Figs. 8(a) and 8(b). When the encoding capacity represented via $I(\mathbf{z};\mathbf{s})$ is low, the optimal strategy is to encode the signal only along the most informative $\hat{\mathbf{w}}_1$ direction. This is reflected in $P(\mathbf{s}|\mathbf{z})$ ellipses being located near the origin and compressed along $\hat{\mathbf{w}}_1$ but not $\hat{\mathbf{w}}_2$ [Fig. 8(a)], and similarly, $P(\mathbf{y}|\mathbf{z})$ ellipses near the origin being compressed along $\mathbf{v}_1$ but not $\mathbf{v}_2$ [Fig. 8(b)]. Thus, for low values of $I(\mathbf{z};\mathbf{s})$, the optimal noise assignment for the second encoding component is $\sigma_2 \to \infty$ [implying $I(z_2;\mathbf{s}) = 0$], while the encoding noise for the first component is set by the encoding capacity via $1/\tilde{\sigma}_1^2 = 2^{2I(\mathbf{z};\mathbf{s})} - 1$ [follows from Eq. (6) with $\tilde{\sigma}_1 = \sigma_1/r$ and $I(z_1;\mathbf{s}) = I(\mathbf{z};\mathbf{s})$], with base-2 used for the log.

In the opposite extreme with very large encoded information $I(\mathbf{z};\mathbf{s})$, the optimal strategy corresponds to low values for both σ_1 and σ_2. This is intuitive because such an assignment allows for an accurate decoding of the vector signal $\mathbf{s}$ from the two encoding components z_1 and z_2. In the limit $I(\mathbf{z};\mathbf{s}) \to \infty$, encoding noises σ_1 and σ_2 approach zero (hence $\ell_1, \ell_2 \to 0$), the decoding distribution $P(\mathbf{s}|\mathbf{z})$ becomes localized at the true signal $\mathbf{s}$ [Fig. 8(a)], and the ellipse corresponding to $P(\mathbf{y}|\mathbf{z})$ on the relevance plane reduces to the ellipse of $P(\mathbf{y}|\mathbf{s})$ with semiaxis lengths a and b, as set by the statistics of $\mathbf{s} \to \mathbf{y}$ mapping [Fig. 8(b)]. Notably, in this limit the relevant information $I(\mathbf{z};\mathbf{y})$ reaches its bound set by the correlations between $\mathbf{s}$ and $\mathbf{y}$ variables, namely, $I_{\max}(\mathbf{z};\mathbf{y}) = I(\mathbf{s};\mathbf{y}) = -\log \tilde{a} - \log \tilde{b}$.

Arguably the most interesting aspect of the Gaussian information bottleneck method is the qualitative change in the optimal encoding strategy that occurs at a critical encoding capacity $I^\dagger(\mathbf{z};\mathbf{s})$. As illustrated geometrically in Fig. 8(b), when the encoded information $I(\mathbf{z};\mathbf{s})$ reaches its critical value, the $P(\mathbf{y}|\mathbf{z})$ decoding ellipse on the relevance plane becomes identical in shape to the $P(\mathbf{y}|\mathbf{s})$ ellipse representing the $\mathbf{s} \to \mathbf{y}$ mapping statistics; at this point, the optimal strategy switches from scalar encoding to vector encoding. Using the Λ_1 and Λ_2 notation for the minor and major semiaxis lengths of the $P(\mathbf{y}|\mathbf{z})$ ellipse, respectively, we note that at the transition point $\Lambda_2^\dagger = r$ and write the similarity condition of $P(\mathbf{y}|\mathbf{z})$ and $P(\mathbf{y}|\mathbf{s})$ ellipses simply as

$$\Lambda_1^\dagger / r = a / b. \tag{21}$$

Here, $\Lambda_1^\dagger$ is the minor axis of the $P(\mathbf{y}|\mathbf{z})$ ellipse at the transition point. It is achieved when the corresponding encoding noise is $\sigma_1^\dagger = \sqrt{\frac{r^2 - b^2}{(b/a)^2 - 1}}$ (see Appendix C2 in the Supplemental Material [18]).

Two special cases are of particular interest. When $b \to a$, corresponding to the case where the two encoding directions are equally informative about the relevance variable, we obtain $\Lambda_1^\dagger \to r$ and $\sigma_1^\dagger \to \infty$. This means that the optimal strategy is to simultaneously encode along both directions as soon as $I(\mathbf{z};\mathbf{s})$ becomes nonzero, which is intuitive because neither of the equally informative directions is given priority over the other. The opposite limit with $b \to r$ corresponds to the case where the second direction is completely uninformative. In this case, we find $\Lambda_1^\dagger \to a$ and $\sigma_1^\dagger \to 0$ [i.e., $I^\dagger(\mathbf{z};\mathbf{s}) \to \infty$], indicating that the optimal strategy is to encode only along the most informative direction $\hat{\mathbf{w}}_1$ for all values of the encoded information $I(\mathbf{z};\mathbf{s})$.

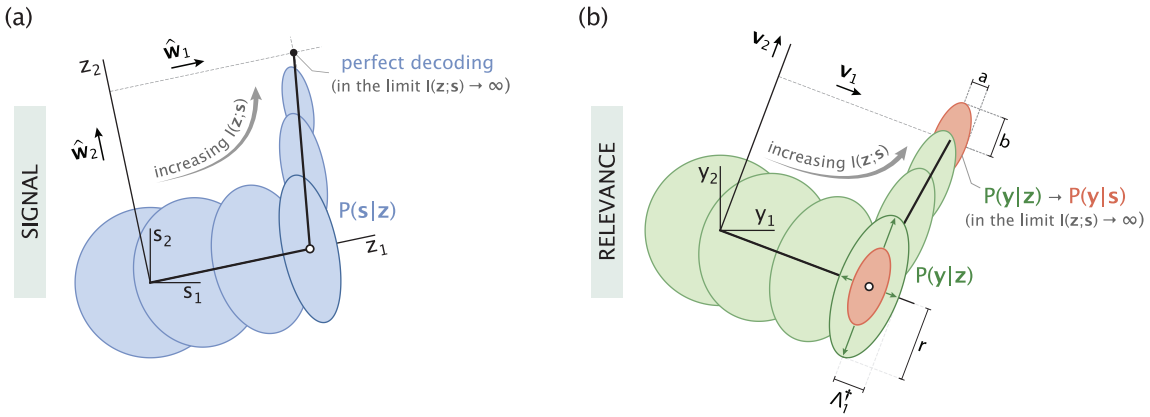

FIG. 8. Geometric illustration of the increasing representation complexity under the optimal strategy. Ellipses corresponding to distributions $P(\mathbf{s}|\mathbf{z})$ and $P(\mathbf{y}|\mathbf{z})$ under optimal encoding [panels (a) and (b), respectively] are shown for different levels of the encoded information $I(\mathbf{z};\mathbf{s})$, ranging from very low to very high. Each pair of corresponding ellipses in panels (a) and (b) is the optimal pair among the family of options for given $I(\mathbf{z};\mathbf{s})$ shown in Fig. 7, maximizing $I(\mathbf{z};\mathbf{y})$ for that given $I(\mathbf{z};\mathbf{s})$. The solid line segments represent the mean vectors $\langle\mathbf{s}|\mathbf{z}\rangle$ and $\langle\mathbf{y}|\mathbf{z}\rangle$. The white dots mark the transition points from a scalar to two-dimensional optimal representation. When tuning $I(\mathbf{z};\mathbf{s})$, the encoding values are kept unchanged at $z_1 = \hat{\mathbf{w}}_1 \cdot \mathbf{s}$ and $z_2 = \hat{\mathbf{w}}_2 \cdot \mathbf{s}$, where $\mathbf{s}$ is the true signal getting encoded.

In addition to offering a simple geometric interpretation of the transition point, the relation in Eq. (21) also motivates an information-theoretic explanation. Recalling that the information contained in the encoding component z_1 about the relevance variable is $I(z_1;\mathbf{y}) = \log(r/\Lambda_1)$, we can write the condition of transition in Eq. (21) in an alternative form as

$$I^\dagger(z_1;\mathbf{y}) = I_a - I_b, \tag{22}$$

where we introduced $I_a = -\log\tilde{a}$ and $I_b = -\log\tilde{b}$. Discussed in the context of Fig. 4, $I_a = I_{\max}(z_1;\mathbf{y})$ and $I_b = I_{\max}(z_2;\mathbf{y})$ are the maximum amounts of relevant information that the encoding components z_1 and z_2, respectively, can contain in the limit of noiseless encoding. Using the $I_{\max}$ notation for I_a and I_b, we rearrange the terms in Eq. (22) and rewrite it as

$$I_{\max}(z_1;\mathbf{y}) - I^\dagger(z_1;\mathbf{y}) = I_{\max}(z_2;\mathbf{y}). \tag{23}$$

This illuminating form suggests the following interpretation of the transition point: the optimal strategy transitions from scalar to vector encoding when the amount of relevant information that can still be stored in the most informative encoding component, namely, $I_{\max}(z_1;\mathbf{y}) - I^\dagger(z_1;\mathbf{y})$, becomes equal to the maximum relevant information $I_{\max}(z_2;\mathbf{y})$ available to the yet unused, less informative encoding component (see Fig. 9).

Referring back to the geometric picture in Fig. 8(b), we now discuss what happens past the transition point as the encoding capacity $I(\mathbf{z};\mathbf{s})$ is increased further. After being established at the transition point, the property of $P(\mathbf{y}|\mathbf{z})$ and $P(\mathbf{y}|\mathbf{s})$ ellipses having an identical aspect ratio persists, i.e., $\Lambda_1/\Lambda_2 = a/b$ when $I(\mathbf{z};\mathbf{s}) > I^\dagger(\mathbf{z};\mathbf{s})$. A similar property also holds in the $\mathbf{s}$ plane [Fig. 8(a)] where the aspect ratio of the decoding $P(\mathbf{s}|\mathbf{z})$ ellipse (generally different from that of the $P(\mathbf{y}|\mathbf{z})$ ellipse) remains unchanged as the $P(\mathbf{s}|\mathbf{z})$ distribution becomes more and more localized with increasing $I(\mathbf{z};\mathbf{s})$. These properties are achieved through a special assignment of encoding noises σ_1 and σ_2 that obey the relations $1/\tilde{\sigma}_1^2 = 2^{I(\mathbf{z};\mathbf{s})+I^\dagger(\mathbf{z};\mathbf{s})} - 1$ and $1/\tilde{\sigma}_2^2 = 2^{I(\mathbf{z};\mathbf{s})-I^\dagger(\mathbf{z};\mathbf{s})} - 1$ (for details, see Appendix C3 in the Supplemental Material [18]).

Additional insights about the behavior beyond the transition point can be obtained by looking at information measures. Specifically, following the approach in Eq. (23), the condition $\Lambda_1/\Lambda_2 = a/b$ can be recast in terms of relevant information amounts as

$$I_{\max}(z_1;\mathbf{y}) - I(z_1;\mathbf{y}) = I_{\max}(z_2;\mathbf{y}) - I(z_2;\mathbf{y}). \tag{24}$$

One can also show that the componentwise encoded information amounts are related via

$$I(z_1;\mathbf{s}) - I^\dagger(z_1;\mathbf{s}) = I(z_2;\mathbf{s}). \tag{25}$$

These relations capture two important behavioral properties of the optimal strategy beyond the transition point. First, the additional encoded information, namely, $I(\mathbf{z};\mathbf{s}) - I^\dagger(\mathbf{z};\mathbf{s})$, is distributed evenly among the two components [Eq. (25)]; second, once equalized at the transition point, the amounts of relevant information that can still be stored in the two encoding components continue to be equal [Eq. (24)]. These properties are captured on the information plane in Fig. 9. In particular, the segment of the $\hat{\mathbf{w}}_1$ curve after the transition point is, except for an offset, identical to the entire $\hat{\mathbf{w}}_2$ curve for the second component; furthermore, the two curves are traversed identically as $I(\mathbf{z};\mathbf{s})$ continues to increase past its critical value $I^\dagger(\mathbf{z};\mathbf{s})$, demonstrating the even allocation of additional encoding capacity and the equality of relevant information amounts that can still be stored in the two components.

V. HIGHER-DIMENSIONAL CASE

Having studied the principles of optimality in the case where the signal and relevance variables are two-dimensional vectors, we proceed in this section to generalizing these principles to the three-dimensional case and, by extension, to an arbitrary multidimensional scenario.

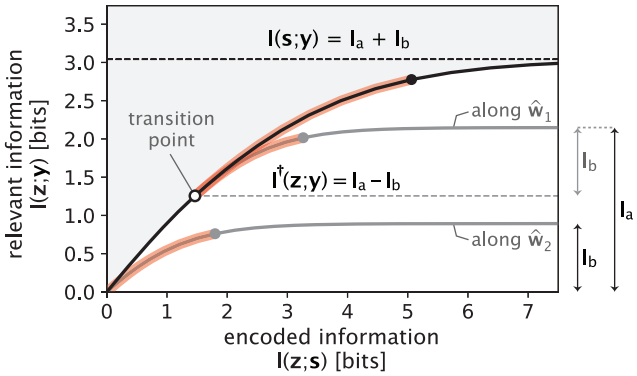

FIG. 9. Increase in encoding complexity and behavior past the transition point illustrated on the information plane. The second encoding component gets introduced when the amount of relevant information that can still be stored in the first component, $I_a - I^{\dagger}(z_1;\mathbf{y})$, becomes equal to the maximum relevant information available to the second component, I_b [Eq. (23)]. The additional encoding capacity past the transition point is allocated equally between the two encoding components. This property is captured through the highlighted segments that indicate the correspondence between a point $[I(\mathbf{z};\mathbf{s}), I(\mathbf{z};\mathbf{y})]$ on the information bound past the transition point (black dot) and the locations $[I(z_i;\mathbf{s}), I(z_i;\mathbf{y})]$ $(i = 1, 2)$ on the corresponding single-component curves (two gray dots). The inaccessible region of the information plane beyond the information bound is colored in gray.

The principal directions of encoding, like in the 2D case, are specified by the eigenvectors of the conditional covariance matrix $\mathbf{\Sigma}_{\mathbf{s}|\mathbf{y}}$. When the variables $\mathbf{s}$ and $\mathbf{y}$ are three dimensional, the three eigenvectors $\hat{\mathbf{w}}_1$, $\hat{\mathbf{w}}_2$, and $\hat{\mathbf{w}}_3$ are parallel to the three axes of the $\mathbf{y} \to \mathbf{s}$ mapping ellipsoid, analogous to the 2D case where $\hat{\mathbf{w}}_1$ and $\hat{\mathbf{w}}_2$ are parallel to the axes of the $\mathbf{y} \to \mathbf{s}$ mapping ellipse [Fig. 3(b)]. We denote the eigenvalues corresponding to these three eigenvectors as a^2, b^2, and c^2, where a, b, and c represent the half-lengths of the axes of the $P(\mathbf{s}|\mathbf{y})$ ellipsoid (with $a \leqslant b \leqslant c$).

Due to the initial standardization performed on $\mathbf{s}$ and $\mathbf{y}$ variables, the $\mathbf{s} \to \mathbf{y}$ mapping ellipsoid in the relevance space has shape and dimensions that are identical to that of the $\mathbf{y} \to \mathbf{s}$ mapping ellipsoid in the signal space [see Fig. 2(b) for the analog in two dimensions]. Therefore, the three axes of the $P(\mathbf{y}|\mathbf{s})$ ellipsoid in relevance space have the same half-lengths, namely, a, b, and c (Fig. 10, left). These three parameters set the mutual information between signal and relevance variables via

$$I(\mathbf{s};\mathbf{y}) = \log\left(\frac{r^3}{a \times b \times c}\right) = -\log\tilde{a} - \log\tilde{b} - \log\tilde{c}, \tag{26}$$

where $\tilde{a} = a/r$ (similarly for $\tilde{b}$ and $\tilde{c}$). From a geometric point of view, $I(\mathbf{s};\mathbf{y})$ is the logarithm of the volume ratio of the marginal $P(\mathbf{y})$ sphere with radius r and the more localized $P(\mathbf{y}|\mathbf{s})$ ellipsoid representing $\mathbf{s} \to \mathbf{y}$ mapping statistics.

Now, the component $z_i = \hat{\mathbf{w}}_i \cdot \mathbf{s} + \xi_i$ that encodes the signal along the direction $\hat{\mathbf{w}}_i$ preserves information about the $\mathbf{y}$ variable along the vector $\mathbf{v}_i = \tilde{\mathbf{\Sigma}}_{\mathbf{ys}}\hat{\mathbf{w}}_i$ in the relevance space. The three vectors $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ corresponding to the set of principal encoding directions $\{\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \hat{\mathbf{w}}_3\}$ described above are perpendicular to one another and are oriented along the axes of the $\mathbf{s} \to \mathbf{y}$ mapping ellipsoid, with $\mathbf{v}_1$ along the shortest axis and $\mathbf{v}_3$ along the longest (Fig. 10, left). How much the encoding values $\{z_i\}$ inform about the relevance variable $\mathbf{y}$ depends on the respective encoding noise strengths $\{\sigma_i^2\}$. To express this dependence, we write the relevant information $I(\mathbf{z};\mathbf{y})$ in the form

$$I(\mathbf{z};\mathbf{y}) = \log\left(\frac{r^3}{\Lambda_1 \times \Lambda_2 \times \Lambda_3}\right) = \underbrace{\log\frac{r}{\Lambda_1}}_{I(z_1;\mathbf{y})} + \underbrace{\log\frac{r}{\Lambda_2}}_{I(z_2;\mathbf{y})} + \underbrace{\log\frac{r}{\Lambda_3}}_{I(z_3;\mathbf{y})}, \tag{27}$$

with

$$\Lambda_i(\sigma_i) = r\sqrt{\frac{\lambda_i^2 + \sigma_i^2}{r^2 + \sigma_i^2}}. \tag{28}$$

Here, $\Lambda_i(\sigma_i)$ is the half-length of $P(\mathbf{y}|\mathbf{z})$ ellipsoid's axis oriented along the vector $\mathbf{v}_i$ and $\lambda_i \in \{a, b, c\}$. In the limit where all three $\sigma_i \to \infty$, no information about the signal is encoded $[I(\mathbf{z};\mathbf{s}) = 0]$, and we find $\Lambda_i \to r$. This is the scenario shown in the left panel of Fig. 10 where the $P(\mathbf{y}|\mathbf{z})$ ellipsoid turns into a sphere corresponding to the marginal $P(\mathbf{y})$. In the opposite limit of noiseless encoding where all three $\sigma_i \to 0$ [hence, $I(\mathbf{z};\mathbf{s}) \to \infty$ and the signal $\mathbf{s}$ can be fully recovered from $\mathbf{z}$], we obtain $\Lambda_i \to \lambda_i$, which means that in this limit, the $P(\mathbf{y}|\mathbf{z})$ ellipsoid will match the $\mathbf{s} \to \mathbf{y}$ mapping ellipsoid of $P(\mathbf{y}|\mathbf{s})$.

To understand the assignment of encoding noise strengths $\{\sigma_i^2\}$ used for optimally navigating between the above two limits, namely, the assignment that maximizes the relevant information $I(\mathbf{z};\mathbf{y})$ for a given finite value of the encoded information

$$I(\mathbf{z};\mathbf{s}) = \sum_{i=1}^{3} \underbrace{\frac{1}{2}\log\left(1 + \frac{r^2}{\sigma_i^2}\right)}_{I(z_i;\mathbf{s})}, \tag{29}$$

we again first consider a geometric perspective to the solution of this problem (Fig. 10). When the encoded information $I(\mathbf{z};\mathbf{s})$ is increased from zero, the $P(\mathbf{y}|\mathbf{z})$ ellipsoid initially gets compressed along one direction only, namely, the direction $\mathbf{v}_1$. This corresponds to the optimal strategy of the scalar encoding along the first principal direction $\hat{\mathbf{w}}_1$. At a critical point where the shape of the $P(\mathbf{y}|\mathbf{z})$ ellipsoid projected onto the $\mathbf{v}_1$-$\mathbf{v}_2$ plane (highlighted in yellow) becomes identical to the shape of the $P(\mathbf{y}|\mathbf{s})$ ellipsoid on that plane, a transition happens where the second component is introduced. The condition of shape identity at the first transition is the same as that in the two-dimensional case [Eq. (21)], namely, $\Lambda_1^{\dagger}/r = a/b$. As the encoded information is increased further, the $P(\mathbf{y}|\mathbf{z})$ ellipsoid gets compressed along two directions, namely, simultaneously along both $\mathbf{v}_1$ and $\mathbf{v}_2$, while keeping its shape on the $\mathbf{v}_1$-$\mathbf{v}_2$ plane constant, i.e., $\Lambda_1/\Lambda_2 = a/b$ after the first transition. When the shape of the $P(\mathbf{y}|\mathbf{z})$ ellipsoid becomes identical to that of $P(\mathbf{y}|\mathbf{s})$ not only on the $\mathbf{v}_1$-$\mathbf{v}_2$ plane, but also on the $\mathbf{v}_2$-$\mathbf{v}_3$ plane, the second transition takes place where now it becomes optimal to have a third encoding component (Fig. 10, right panel). The similarity condition of $P(\mathbf{y}|\mathbf{z})$ and $P(\mathbf{y}|\mathbf{s})$

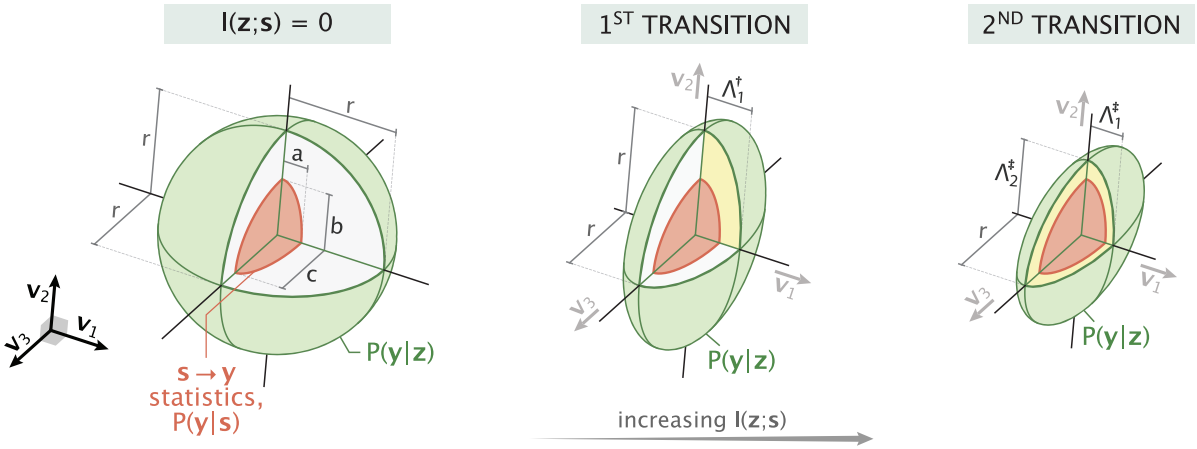

FIG. 10. Geometric perspective on the transitions in representation complexity with growing encoding capacity in a three-dimensional setting. The red ellipsoid corresponding to the $\mathbf{s} \to \mathbf{y}$ mapping statistics (shown only in the first octant) has dimensions that are independent of $I(\mathbf{z};\mathbf{s})$. The three axes of this ellipsoid are oriented along the perpendicular vectors $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$. When no information is encoded, the $P(\mathbf{y}|\mathbf{z})$ ellipsoid takes the spherical shape of the marginal $P(\mathbf{y})$. As $I(\mathbf{z};\mathbf{s})$ increases from zero, the encoding component z_1 retains information about the relevance variable $\mathbf{y}$ along the direction $\mathbf{v}_1$, and the $P(\mathbf{y}|\mathbf{z})$ ellipsoid gets compressed along that direction. At a critical value $I^\dagger(\mathbf{z};\mathbf{s})$, the shapes of $P(\mathbf{y}|\mathbf{z})$ and $P(\mathbf{y}|\mathbf{s})$ ellipsoids projected onto the $\mathbf{v}_1$-$\mathbf{v}_2$ plane (highlighted in yellow) become identical. At that point, the second encoding component z_2 is introduced that retains information about $\mathbf{y}$ along the direction $\mathbf{v}_2$. After shrinking simultaneously along both $\mathbf{v}_1$ and $\mathbf{v}_2$, at the second critical value $I^\ddagger(\mathbf{z};\mathbf{s})$, the $P(\mathbf{y}|\mathbf{z})$ ellipsoid becomes identical in shape to $P(\mathbf{y}|\mathbf{s})$ also on the $\mathbf{v}_2$-$\mathbf{v}_3$ plane. At that point, the third encoding component z_3 gets introduced, which informs on $\mathbf{y}$ along the direction $\mathbf{v}_3$. In the limit $I(\mathbf{z};\mathbf{s}) \to \infty$, where the signal $\mathbf{s}$ can be fully recovered from $\mathbf{z}$, the $P(\mathbf{y}|\mathbf{z})$ and $P(\mathbf{y}|\mathbf{s})$ ellipsoids become identical.

ellipsoids at the second transition is

$$\frac{\Lambda_2^\ddagger}{r} = \frac{b}{c}. \tag{30}$$

Past this transition, $P(\mathbf{y}|\mathbf{z})$ gets compressed along all three directions proportionally, i.e.,

$$\Lambda_1 \div \Lambda_2 \div \Lambda_3 = a \div b \div c, \tag{31}$$

until the distributions $P(\mathbf{y}|\mathbf{z})$ and $P(\mathbf{y}|\mathbf{s})$ and their corresponding shapes fully overlap in the limit $I(\mathbf{z};\mathbf{s}) \to \infty$.

Next, we discuss how these features of the optimal strategy get reflected on the information plane [Fig. 11(a)]. As in the two-dimensional scenario considered earlier (Fig. 9), the transition from scalar to 2D optimal encoding happens when the amount of relevant information that can still be stored in the first component, namely, $I_a - I^\dagger(z_1;\mathbf{y})$, becomes equal to the maximum amount that can be stored in the second component, I_b.

To deduce the condition for the second transition (2D → 3D optimal encoding), we recast Eq. (30) in terms of information metrics as

$$I_b - I^\ddagger(z_2;\mathbf{y}) = I_c. \tag{32}$$

Here, we used the identities $I_b = -\log \tilde{b}$, $I_c = -\log \tilde{c}$, and $I^\ddagger(z_2;\mathbf{y}) = \log(r/\Lambda_2^\ddagger)$. The second transition thus occurs when the amount of relevant information that z_2 can still store, $I_b - I^\ddagger(z_2;\mathbf{y})$, which is the same as the amount of relevant information that z_1 can still store, $I_a - I^\ddagger(z_1;\mathbf{y})$, becomes equal to the maximum amount that can be stored in the third component, namely, I_c. The total relevant information at the second transition is then equal to

$$\begin{aligned} I^\ddagger(\mathbf{z};\mathbf{y}) &= I^\ddagger(z_1;\mathbf{y}) + I^\ddagger(z_2;\mathbf{y}) \\ &= I_a + I_b - 2I_c. \end{aligned} \tag{33}$$

After the second transition, as in the two-dimensional case, any additional encoded information is allocated equally among the three components, yielding identical amounts of additional relevant information from each of the components.

The information-theoretic principles behind the increasing representation complexity and optimal behavior between distinct transitions can be seen even more directly in the componentwise information plots shown in Fig. 11(b). There, the relevant information that can still be stored in component i, namely, $I_{\max}(z_i;\mathbf{y}) - I(z_i;\mathbf{y})$, is plotted against the amount of encoded information allocated to that component, $I(z_i;\mathbf{s})$. As indicated by the arrows, the curves are traversed top-down, with a given horizontal level corresponding to a location on the information bound in Fig. 11(a), which is uniquely specified by the Lagrange multiplier γ that sets the derivative via $\gamma = \partial I(\mathbf{z};\mathbf{y})/\partial I(\mathbf{z};\mathbf{s})$. In fact, this derivative is the same as that along the componentwise curves that together give rise to the information bound, i.e., $\gamma = \partial I(z_i;\mathbf{y})/\partial I(z_i;\mathbf{s})$ after the ith component is introduced. As γ decreases, new components get introduced at special moments where the horizontal level, indicating the amount of relevant information that can still be stored in each of the existing components, matches the maximum level of the new component [Fig. 11(b)]. Once introduced, the curve of the new component is traversed identically to the curves of existing components.

In fact, one can also draw a thermodynamic analogy between the optimal allocation of encoded information across different z components and the problem of optimally distributing particles among boxes of varying sizes. Specifically, the

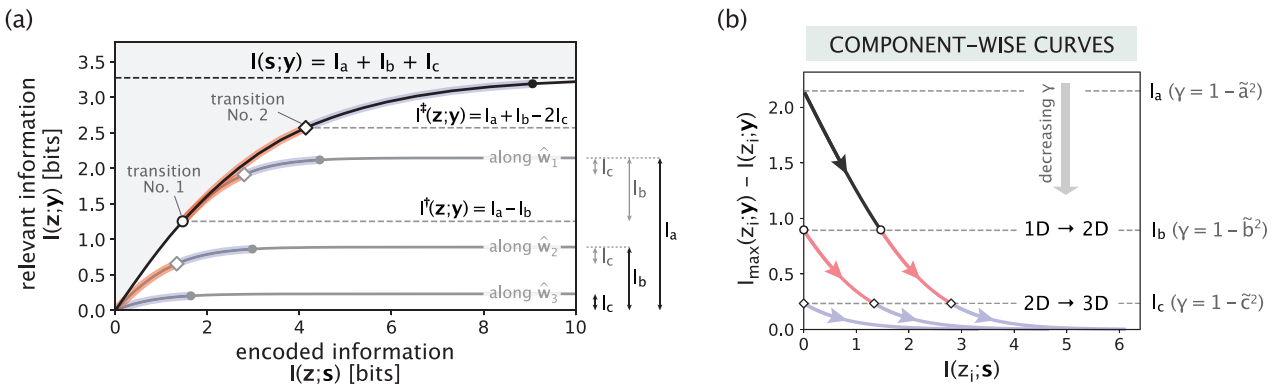

FIG. 11. Increasing representation complexity illustrated on the information plane in the case where signal and relevance variables are three dimensional. (a) Piecewise formation of the information bound $I(\mathbf{z};\mathbf{y})$ vs $I(\mathbf{z};\mathbf{s})$. Before the first transition, only the most informative first component (encoding along $\hat{\mathbf{w}}_1$) is used. After the first and before the second transition, components encoding along $\hat{\mathbf{w}}_1$ and $\hat{\mathbf{w}}_2$ are used simultaneously (highlighted in red). The second transition occurs when the relevant information that can still be stored in each of the first two components becomes equal to the maximum amount (I_c) that can be stored in the third component encoding along $\hat{\mathbf{w}}_3$. Beyond the second transition, the three single-component curves are traversed identically, yielding the blue highlighted segment on the information bound. The three gray dots at the ends of single-component curves with coordinates $[I(z_i;\mathbf{s}), I(z_i;\mathbf{y})]$ ($i \in \{1,2,3\}$) correspond to the black dot on the information bound via $I(\mathbf{z};\mathbf{s}) = \sum_i I(z_i;\mathbf{s})$ and $I(\mathbf{z};\mathbf{y}) = \sum_i I(z_i;\mathbf{y})$. (b) Componentwise plots of the relevant information that can still be stored in a component vs the information encoded in that component ($i = 1, 2, 3$ correspond to the three respective curves, from top to bottom). Arrows indicate the top-down traversal of the curves, with a given horizontal level corresponding to a unique location on the information bound specified by the Lagrange multiplier γ. The black, red, and blue colors of curve segments indicate scalar, 2D, and 3D optimal encoding regimes, respectively.

boxes of varying sizes represent the different amounts of maximum relevant information that the z components can store. Placing a particle in a given box corresponds to allocating additional encoded information to a given z component. Thermodynamically, the optimal distribution of particles in the boxes should be such that the total free energy is minimized. In general, the optimal strategy is to initially fill the larger box and begin filling the smaller box only when the chemical potentials of particles in the two boxes become equal. One can also show that if the particles are distinguishable, the second box starts to fill when the remaining available volume in the first box equals that of the smaller second box. From that point onward, both boxes should be filled simultaneously, maintaining equal available volumes and chemical potentials. The same principle extends to the case of more than two boxes (see Appendix D3 in the Supplemental Material [18] for details). This is analogous to how in the optimal encoding strategy it is best to use the most informative component first until the relevant information that can still be stored in that component equals the maximum amount that can be stored in the second, yet unused component.

Taken together, the diverse perspectives presented above offer an intuitive understanding of the various aspects of the optimal encoding strategy—from the optimal choice of encoding directions, to the allocation of encoding capacity among these directions, to the emergence of distinct transitions in the complexity of signal representation.

A. General high-dimensional case

The intuition and principles demonstrated earlier for the two- and three-dimensional cases extend naturally to arbitrary dimensions. Reserving the detailed discussion of the general solution to the bottleneck problem to Appendix D1 in the Supplemental Material [18], here we present some of its main results that go in parallel to the results of Chechik *et al.* [10]. In particular, the noise strength associated with the ith encoding direction can be expressed as

$$\sigma_i^2(\gamma) = r^2 \frac{\gamma \tilde{\lambda}_i^2}{(1-\gamma) - \tilde{\lambda}_i^2}, \tag{34}$$

where γ is the Lagrange multiplier, r is the scale of the standardized signal ($\mathbf{\Sigma}_\mathbf{s} = r^2 \mathbf{I}_\mathbf{s}$), and λ_i is the ith semiaxis length of the $P(\mathbf{s}|\mathbf{y})$ ellipsoid (with $\tilde{\lambda}_i = \lambda_i / r$); equivalently, λ_i^2 is the ith eigenvalue of the conditional covariance matrix $\mathbf{\Sigma}_{\mathbf{s}|\mathbf{y}}$. The ith encoding component is introduced when the value of the Lagrange multiplier gets below the critical value

$$\gamma_i^c = 1 - \tilde{\lambda}_i^2. \tag{35}$$

With $\sigma_i^2(\gamma)$ at hand, we next compute the ith semiaxis length $\ell_i(\gamma)$ of the decoding $P(\mathbf{s}|\mathbf{z})$ ellipsoid, namely,

$$\ell_i(\gamma) = r \sqrt{\frac{\gamma}{1-\gamma}} \sqrt{\frac{\tilde{\lambda}_i^2}{1-\tilde{\lambda}_i^2}}. \tag{36}$$

The fact that the γ dependence appears as a multiplicative factor indicates that the aspect ratio $\ell_i(\gamma) \div \ell_j(\gamma)$ stays independent of γ once the ith and jth components are both introduced.

Similarly, we can compute the ith semiaxis length $\Lambda_i(\gamma)$ of the $P(\mathbf{y}|\mathbf{z})$ ellipsoid in the relevance space as

$$\Lambda_i(\gamma) = \frac{\lambda_i}{\sqrt{1-\gamma}}. \tag{37}$$

One can easily see how the expression for the aspect ratio follows, namely, $\Lambda_i(\gamma) \div \Lambda_j(\gamma) = \lambda_i \div \lambda_j$. This shows the generality of the geometric interpretation discussed earlier in cases of 2D [Fig. 8] and 3D [Fig. 10] encoding.

Finally, we note that the encoded information $I(\mathbf{z};\mathbf{s})$ and the relevant information $I(\mathbf{z};\mathbf{y})$ can be obtained directly from the semiaxis length expressions via

$$I(\mathbf{z};\mathbf{s}) = \log \prod_{i=1}^{n(\gamma)} \frac{r}{\ell_i(\gamma)} = -\sum_{i=1}^{n(\gamma)} \log \tilde{\ell}_i(\gamma), \tag{38}$$

$$I(\mathbf{z};\mathbf{y}) = \log \prod_{i=1}^{n(\gamma)} \frac{r}{\Lambda_i(\gamma)} = -\sum_{i=1}^{n(\gamma)} \log \tilde{\Lambda}_i(\gamma), \tag{39}$$

where $n(\gamma)$ is the number of encoding components that have already been introduced for the given value of the Lagrange multiplier γ.

B. Comment on the degenerate space of optimal solutions

In our general treatment of the bottleneck problem so far, we considered the encoding components $z_i = \hat{\mathbf{w}}_i \cdot \mathbf{s} + \xi_i$ to be statistically independent of one another. This was achieved by choosing independent encoding noises (ξ_i) and perpendicular encoding directions ($\hat{\mathbf{w}}_i$) corresponding to the eigenvectors of $\mathbf{\Sigma}_{\mathbf{s}|\mathbf{y}}$. A broader space of optimal solutions, however, becomes accessible when relaxing the assumption of independent components. Specifically, the same $I(\mathbf{z};\mathbf{s})$ and $I(\mathbf{z};\mathbf{y})$ can be achieved by considering correlated noises and/or nonperpendicular encoding directions that correlate z_i values through the signal $\mathbf{s}$.

A particularly interesting class of degenerate optimal solutions is the one where the encoding noises assigned to nonperpendicular directions are uncorrelated but have the same strength. This is in contrast to the principal solution where the more informative components have a lower associated noise (hence, a higher allocated encoding capacity). The identical noise strength in such alternative scheme, given by $\sigma_{\text{alt}}^{-2} = \langle \sigma_i^{-2} \rangle$ (averaging performed over different components i), falls between the highest and the lowest ones assigned to the principal directions, i.e., $\sigma_1 \leqslant \sigma_{\text{alt}} \leqslant \sigma_n$, where n is the number of components. An analogous inequality then holds for individual encoded information amounts, namely, $I(z_1;\mathbf{s}) \geqslant I(z_i^{\text{alt}};\mathbf{s}) \geqslant I(z_n;\mathbf{s})$, where $I(z_i^{\text{alt}};\mathbf{s})$ is the same for all i because σ_{alt} is the same. While $I(\mathbf{z}^{\text{alt}};\mathbf{s}) = I(\mathbf{z};\mathbf{s})$ from the optimality condition, due to the correlations between z_i^{alt}, the sum of componentwise information amounts in the alternative strategy will be greater than that for the principal strategy, namely,

$$\underbrace{\sum_{i=1}^{n} I(z_i^{\text{alt}};\mathbf{s})}_{nI(z_i^{\text{alt}};\mathbf{s})} \geqslant \underbrace{\sum_{i=1}^{n} I(z_i;\mathbf{s})}_{I(\mathbf{z};\mathbf{s})}. \tag{40}$$

The relation in Eq. (40) may have functional implications in terms of encoding cost. For example, if the energetic or resource cost of encoding the signal into different components were to scale linearly with the sum of the individual information amounts, then the principal solution with perpendicular encoding directions would be the preferred one. In the principal solution, however, encoding capacity is distributed unequally across components (i.e., $I(z_1;\mathbf{s}) \geqslant I(z_2;\mathbf{s}) \geqslant \cdots$). This means that if the encoding cost were to scale nonlinearly with information [8,20] (see also Refs. [21,22] for nonlinear optimization Lagrangians), then the alternative strategy with nonperpendicular directions may become the preferred one. We refer the reader to Appendix D2 in the Supplemental Material [18] for a more detailed discussion of the degenerate space of solutions.

VI. SIGNAL PREDICTION PROBLEM

As an application of the different perspectives on the bottleneck method, we now examine the problem of signal prediction for a canonical signal-generation model [3,7,8,16,17]. In particular, we consider signals arising from a stochastically driven harmonic oscillator, the dynamics of which is governed by the following set of dimensionless equations:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = v, \tag{41}$$

$$\frac{\mathrm{d}v}{\mathrm{d}t} = -x - \eta v + \sqrt{2\eta}\,\psi(t). \tag{42}$$

Here, the position x represents the signal, v is the signal derivative, $\psi(t)$ denotes unit white noise, and η is the damping coefficient. The value of η dictates the qualitative behavior of signal dynamics: underdamped for $\eta < 2$, critically damped at $\eta = 2$, and overdamped when $\eta > 2$.

In the context of information bottleneck, the prediction problem is to encode the past signal trajectory in such a way that the encoding preserves maximum information about the future trajectory, subject to a constraint on encoding capacity. Importantly, while the dynamics of x is non-Markovian, the joint dynamics of x and v is Markovian, as per Eqs. (41) and (42). As a result, predicting the future trajectory $x_{[\tau,\infty)}$, where τ is the forecast interval, reduces to predicting the pair (x_τ, v_τ). Similarly, out of the entire past trajectory $x_{(-\infty,0]}$, only the present value and its derivative, namely, (x_0, v_0), need to be encoded for the purpose of predicting (x_τ, v_τ).

The problem of optimal encoding can then be framed in the two-dimensional setting studied earlier in Sec. IV. Specifically, the goal is to optimally encode the current position x_0 and derivative v_0 into the variables

$$z_1 = \hat{w}_{1,1}\, x_0 + \hat{w}_{1,2}\, v_0 + \xi_1, \tag{43a}$$

$$z_2 = \hat{w}_{2,1}\, x_0 + \hat{w}_{2,2}\, v_0 + \xi_2, \tag{43b}$$

such that $\mathbf{z} = [z_1, z_2]^{\mathrm{T}}$ is maximally informative about the future position x_τ and derivative v_τ, subject to a constraint on the encoded information $I(\mathbf{z};\mathbf{s}_0)$. Here, we introduced the state vector $\mathbf{s}_0 = [x_0, v_0]^{\mathrm{T}}$ to represent the initial pair (x_0, v_0). Similarly, we will use $\mathbf{s}_\tau = [x_\tau, v_\tau]^{\mathrm{T}}$ to denote the pair at time τ, which, in the information bottleneck terminology, will be the relevance variable $\mathbf{y}$.

We note that this problem setup was also studied in the prior work of Sachdeva *et al.* [7]. Our approach, however, differs in several key aspects. First, the marginal distributions $P(\mathbf{s}_0)$ and $P(\mathbf{s}_\tau)$ resulting from the dimensionless equations of dynamics [Eqs. (41) and (42)] are bivariate Gaussians with covariance matrices equal to the identity matrix, i.e., $\langle x^2 \rangle = \langle v^2 \rangle = 1$ and $\langle xv \rangle = 0$ (see Appendixes E1 and E2 in the Supplemental Material [18]). With this standardized form,

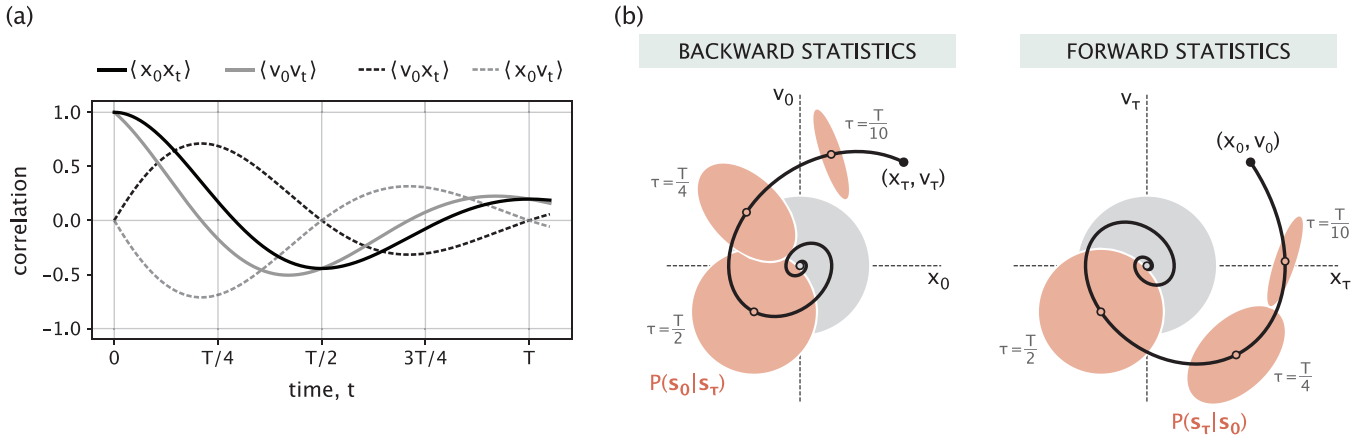

FIG. 12. Statistical structure of an example underdamped signal with $\eta = 0.5$. (a) Autocorrelation and cross-correlation functions of the signal and its derivative. (b) Backward and forward signal statistics represented by the distribution $P(\mathbf{s}_0|\mathbf{s}_\tau)$ and $P(\mathbf{s}_\tau|\mathbf{s}_0)$, respectively, for difference values of τ. Backward statistics capture the distribution of present states $\mathbf{s}_0$ that could lead to the given future state $\mathbf{s}_\tau$, whereas forward statistics capture the distribution of future states $\mathbf{s}_\tau$ that may be reached from the given present state $\mathbf{s}_0$. Gray circles of unit radius correspond to the marginal $P(\mathbf{s})$ reached in the $\tau \to \infty$ limit.

the orientations of encoding vectors $\hat{\mathbf{w}}_1$ and $\hat{\mathbf{w}}_2$ directly reflect the relative importance assigned to x_0 and v_0 during encoding. In the treatment by Sachdeva *et al.* [7], however, a different nondimensionalization approach leads to unequal marginal variances for x_0 and v_0. As a result, the weights they derive are not direct indicators of the relative importance assigned to x_0 and v_0, and the corresponding analytical expressions are more complex and harder to interpret. Second, Sachdeva *et al.* [7] focus on the study of the leading encoding component z_1, addressing the second component z_2 only briefly and in general terms. In this work, we place equal emphasis on the second component and its predictive capabilities across varying forecast intervals, uncovering diverse behaviors.

We start off with the question of finding the optimal encoding directions $\hat{\mathbf{w}}_1$ and $\hat{\mathbf{w}}_2$. These directions are determined by the future $(\mathbf{s}_\tau) \to$ present $(\mathbf{s}_0)$ mapping statistics. In turn, these statistics—captured by the conditional covariance matrix $\mathbf{\Sigma}_{\mathbf{s}_0|\mathbf{s}_\tau}$—are set by the pairwise autocorrelation and cross-correlation functions of position and velocity. As derived in Appendix E3 in the Supplemental Material [18], the pairwise correlations determine the orientation angles of $\hat{\mathbf{w}}_1$ and $\hat{\mathbf{w}}_2$ via

$$\tan(2\varphi) = \frac{2\langle v_0 x_\tau\rangle}{\langle x_0 x_\tau\rangle + \langle v_0 v_\tau\rangle}. \tag{44}$$

This condition is simultaneously satisfied by the angle φ_1 of $\hat{\mathbf{w}}_1$ and the angle $\varphi_2 = \varphi_1 + \pi/2$ of the perpendicular direction $\hat{\mathbf{w}}_2$.

Example correlation functions for underdamped dynamics with $\eta = 0.5$ are shown in Fig. 12(a). There, as can be seen, the correlation terms exhibit damped oscillations, each with its own phase. The backward and forward signal statistics derived from these pairwise correlations are illustrated in Fig. 12(b) for different choices of the forecast interval τ. The ellipses corresponding to these statistics increase in area with increasing forecast interval τ, eventually converging to the circle of the marginal distribution (shown in gray) as $\tau \to \infty$.

The optimal encoding angles φ_1 and φ_2 correspond to the orientation angles of the minor and major axes, respectively, of the backward statistics ellipses [Fig. 12(b)]. One can show that, irrespective of the damping regime or the forecast interval, the angle φ_1 always lies in the first quadrant. This is reflected in the counterclockwise tilt of the backward statistics ellipses [Fig. 12(b)]. Thus, the first encoding component combines the current signal and its derivative constructively (with weights of the same sign), whereas the second component combines them destructively (with weights of opposite signs). The angle φ_1 for different damping regimes is given by

$$\varphi_1 = \begin{cases} \frac{1}{2}\arctan\left(\frac{\tanh(\kappa\tau)}{\kappa}\right), & \text{if } \eta > 2, \\ \frac{1}{2}\arctan\left(\frac{\tan(\omega\tau)}{\omega}\right), & \text{if } \eta < 2, \\ \frac{1}{2}\arctan\tau, & \text{if } \eta = 2. \end{cases} \tag{45}$$

Here, $\kappa = \sqrt{\eta^2/4 - 1}$ is defined for the overdamped regime $(\eta > 2)$ and $\omega = \sqrt{1 - \eta^2/4}$ is defined for the underdamped regime $(\eta < 2)$. Importantly, we restrict the output of the arctan function to the range $(0, \pi)$, which ensures that the angle φ_1 lies within $(0, \pi/2)$ and thus remains in the first quadrant, while the angle $\varphi_2 = \varphi_1 + \pi/2$ remains in the second quadrant.

We focus our further analysis of the optimal strategy on the underdamped case, which is arguably the more interesting regime, deferring the discussion of overdamped and critically damped cases to Appendix E3 in the Supplemental Material [18]. Figure 13(a) shows the encoding angles φ_1 and $\varphi_2 = \varphi_1 + \pi/2$ as functions of the forecast interval τ in the underdamped regime with $\eta = 0.5$. The nearly linear dependence on τ can be easily deduced from Eq. (45), where $\omega = \sqrt{1 - \eta^2/4} \approx 1$ for $\eta = 0.5$, which results in

$$\varphi_1(\tau) \approx \frac{\tau}{2} \bmod \frac{\pi}{2}. \tag{46}$$

For short forecast intervals, the autocorrelation $\langle x_0 x_\tau\rangle$ decays less than $\langle v_0 v_\tau\rangle$ [Fig. 12(a)]. Thus, for small τ, the dominant

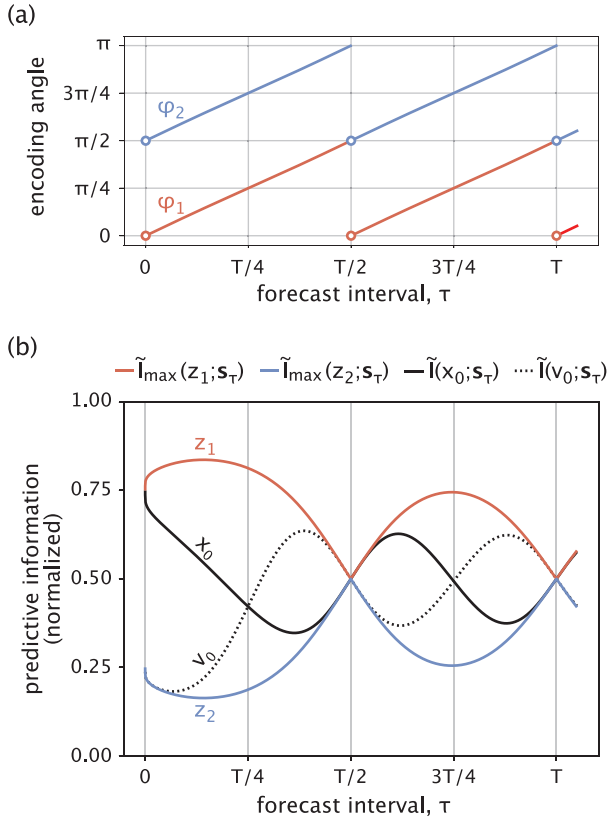

FIG. 13. Features of the optimal encoding strategy for predicting signals arising from underdamped dynamics ($\eta = 0.5$). (a) Optimal encoding angles φ_1 and $\varphi_2 = \varphi_1 + \pi/2$ for the components z_1 and z_2, respectively, shown as a function of the forecast interval τ. The angle φ_1 lies in the range $(0, \pi/2)$, while φ_2 lies in $(\pi/2, \pi)$. (b) Normalized values of predictive information plotted against the forecast interval for different encoding strategies. All strategies are evaluated in the zero-noise limit, $I(z_i; \mathbf{s}_0) \to \infty$. Normalization is performed by the maximum achievable predictive information, namely, $I(\mathbf{s}_0; \mathbf{s}_\tau)$, as determined by the signal statistics. The upper (red) and lower (blue) curves represent the normalized predictive information available to the z_1 and z_2 components, respectively, in the limit of infinite encoding capacity. Their sum equals 1, namely, $\tilde{I}_{\max}(z_1; \mathbf{s}_\tau) + \tilde{I}_{\max}(z_2; \mathbf{s}_\tau) = 1$. The two curves in between correspond to purely x_0-based (solid black line) and purely v_0-based (dotted black line) strategies. Analogous plots for the overdamped regime are shown in Fig. A3 in the Supplemental Material [18].

encoding component z_1 prioritizes the current signal value x_0 over the current derivative v_0. When the forecast interval exceeds a quarter period, $T/4$ (with $T = 2\pi/\omega$), the autocorrelation $\langle v_0 v_\tau \rangle$ exceeds $\langle x_0 x_\tau \rangle$ in magnitude [Fig. 12(a)], and the priority shifts from x_0 to v_0. As τ approaches half a period $T/2$, z_1 becomes predominantly v_0 based. This behavior repeats every half-period. Indeed, as the dominant component, z_1 prioritizes the more informative signal feature at the given forecast interval. The second component z_2, encoding in an orthogonal direction, does the opposite, reversing the prioritization between x_0 and v_0.

Additional insights about the predictive capacity of the two encoding components can be gained by examining the information metrics. Specifically, we consider the maximum predictive information available to z_1 and z_2 components, since the optimal number of encoding dimensions (1D or 2D) at finite encoding capacities depends on how different these maximum values are [e.g., see Fig. 9, with $I_a = I_{\max}(z_1; \mathbf{s}_\tau)$ and $I_b = I_{\max}(z_2; \mathbf{s}_\tau)$]. Figure 13(b) shows the normalized values of these information measures as a function of the forecast interval. Each measure is normalized by the upper bound on predictive information set by the signal statistics, i.e., $I(\mathbf{s}_0; \mathbf{s}_\tau) = I_{\max}(z_1; \mathbf{s}_\tau) + I_{\max}(z_2; \mathbf{s}_\tau)$. The plot thus shows how the relative importance of the two components varies with the forecast interval. In the same figure, we also plot the normalized predictive information values corresponding to purely x_0-based and purely v_0-based strategies.

Figure 13(b) reveals several important aspects of the two components' behavior. As already recognized from Fig. 13(a), the dominant component z_1 is x_0 based and the second component z_2 is v_0 based in the limit $\tau \to 0$. In this limit, the first component always contains 3 times as much predictive information as the second component, irrespective of the damping regime (see Appendix E3 in the Supplemental Material [18]). Interestingly, the strategy based on z_1 significantly outperforms the purely x_0-based strategy already at short forecast intervals. This demonstrates the importance of incorporating the current signal derivative v_0 in the encoding, despite its relatively small weight [$\hat{w}_{1,2} = \sin\varphi_1 \ll 1$ for small τ; see Eqs. (43a) and (46)]. At a forecast interval equal to a quarter period ($\tau = T/4$), x_0 and v_0 become equally predictive of $\mathbf{s}_\tau$ [$\langle x_0 x_\tau \rangle = -\langle v_0 v_\tau \rangle$ and $\langle x_0 v_\tau \rangle = -\langle v_0 x_\tau \rangle$; see Fig. 12(a)], and the optimal strategy combines them with equal weights [$\hat{w}_{1,1} = \hat{w}_{1,2} = 1/\sqrt{2}$ since $\varphi_1(T/4) = \pi/4$]. As τ approaches half a period, z_1 becomes predominantly v_0 based, while z_2 becomes predominantly x_0 based. The predictive capacities of the two components also converge, becoming equal at $\tau = T/2$, as reflected in the circular shape of the backward statistics ellipse at that forecast interval [Fig. 12(b)]. A similar pattern repeats every half-period. In subsequent cycles, the relative predictive capacity of the z_1 component reaches its maximum at $\tau_k \approx T/4 + k(T/2)$, for $k \geqslant 1$. For sufficiently long forecast intervals ($\eta\tau_k \gg 1$), this peak value is given by (see Appendix E3 in the Supplemental Material [18])

$$\tilde{I}_{\max}(z_1; \mathbf{s}_{\tau_k}) \approx \frac{1}{2} + \frac{\eta/2}{1 + (\eta/2)^2}. \tag{47}$$

The monotonic dependence of $\tilde{I}_{\max}(z_1; \mathbf{s}_{\tau_k})$ on the damping coefficient η shows that the gap in the predictive capacities of the z_1 and z_2 components decreases with weaker damping (lower η). Thus, the weaker the damping, the more important it becomes to incorporate the second component for accurate prediction.

Altogether, our reexamination of the signal prediction problem studied earlier by Sachdeva *et al.* [7] has led to a deeper understanding of the encoding features underlying optimal prediction. First, the concise analytical expressions for the optimal encoding angle [Eq. (45)], derived from the standardized signal statistics, provide direct, quantitative insights into how the optimal prioritization of x_0 over v_0 is determined in different settings. Second, our extension of the study to multiple encoding components has revealed an intricate dependence of their predictive capacities on the forecast interval,

which, in turn, varies qualitatively across damping regimes (Fig. 13; see also Fig. A3 in the Supplemental Material [18]).

VII. DISCUSSION

The information bottleneck method is an established framework for learning efficient signal representations that are maximally informative about a relevance variable, subject to limits on the encoding capacity [6]. In a special, yet important case where the signal and relevance variables are jointly Gaussian, the optimal representation is also Gaussian, and an analytical solution to the bottleneck problem becomes available. Since the original work by Chechik *et al.* [10], the Gaussian bottleneck method has been applied to speaker recognition [23], used in studies of cellular prediction [8], explored for its parallels with the renormalization group [24,25], and extended to more general problem settings [26–28]. Despite its diverse applications, our current understanding of the Gaussian information bottleneck method and its emergent features remains largely mathematical.

The main aim of our work was to advance our understanding of the method from multiple, mutually enriching perspectives. The initial standardization of the problem variables enabled us to extract geometric insights into the optimal strategy, without compromising the generality of the analysis. In particular, by studying the less commonly considered decoding distribution $P(\mathbf{y}|\mathbf{z})$ in the relevance space and its relationship to the distribution $P(\mathbf{y}|\mathbf{s})$ of the stochastic $\mathbf{s} \to \mathbf{y}$ mapping, we found distinct signatures of the optimal strategy in the geometric depictions of these distributions (ellipses in 2D, ellipsoids in higher dimensions). First, their axes are aligned, with each aligned pair corresponding to the same rank in the ascending order of lengths [i.e., the shortest axis of $P(\mathbf{y}|\mathbf{z})$ is aligned with the shortest axis of $P(\mathbf{y}|\mathbf{s})$, and so on]. Second, and more importantly, transitions from lower- to higher-dimensional optimal representations occur when the aspect ratios of these axes become equal at critical levels of encoded information (Fig. 10).

The latter geometric insight then hinted at the existence of an information-theoretic criterion for the bifurcation points and the optimal navigation between them. Specifically, an encoding dimension is introduced when the amount of relevant information that each of the existing components can still store becomes equal to the maximum relevant information available to the new, unused component. After the new component is introduced, the additional encoded information is allocated equally among all components, yielding identical amounts of additional relevant information from each component (Fig. 11). Interestingly, this principle is reminiscent of the thermodynamic problem of distributing particles among boxes of different sizes in such a way that the total free energy is minimized. The optimal strategy is to first fill the larger box and only begin filling the smaller box when the chemical potentials of particles in the two boxes become equal.

The canonical solution to the Gaussian information bottleneck problem is expressed in terms of statistically independent encoding components. With the standardization procedure applied in our treatment ($\mathbf{\Sigma}_\mathbf{s} = r^2\mathbf{I}_\mathbf{s}$, $\mathbf{\Sigma}_\mathbf{y} = r^2\mathbf{I}_\mathbf{y}$), these components encode along a set of orthonormal directions (eigenvectors of $\mathbf{\Sigma}_{\mathbf{s}|\mathbf{y}}$) and have independent sources of encoding noise, with the more informative directions assigned lower noise levels or, equivalently, higher encoded information. The space of optimal solutions, however, is degenerate, composed of alternative strategies that involve correlated encoding components. One noteworthy case is when the total encoding capacity is distributed equally among nonorthogonal encoding directions, leading to components correlated via the signal, with each component assigned the same intermediate level of independent noise. If the resource cost of encoding the signal into a component does not scale linearly with the amount of information encoded [8], and/or if the high-precision (low σ) encoding of the dominant component in the canonical setup becomes prohibitive, then an alternative strategy with correlated components may, in fact, be advantageous.

As a practical demonstration of our perspective on the bottleneck method, we next revisited the previously studied problem of predicting signals coming from stochastically driven harmonic oscillator dynamics [3,7,8,16,17]. This problem conveniently reduces to optimally encoding the current signal value and its derivative, as these two features of the past signal trajectory fully specify the future signal statistics. Our use of standardized signal statistics enabled concise analytical expressions for the orientation angles of the $P(\mathbf{s}|\mathbf{y})$ ellipse, which directly indicate the relative weighting of the current signal and its derivative in the two encoding components. Furthermore, our extended analysis of the second encoding component revealed the complex dependence of its predictive capacity on the forecast interval as well as on the damping regime of the signal dynamics. In the future, it would be fruitful to also investigate the energetic requirements of signal prediction by linking the information-theoretic metrics used in the bottleneck method to concepts such as work and heat dissipation used in thermodynamics (see, e.g., Ref. [29]).

The prevalence of linear models with Gaussian statistics in natural and engineered systems makes the Gaussian information bottleneck a practically applicable framework, in addition to serving as an analytically tractable baseline against which more sophisticated numerical frameworks can be compared. Developing a deep, intuitive understanding of its various features is therefore essential—an aim that our work has contributed to.

For clarity of exposition, particularly regarding the geometric interpretations, we focused our analysis on the low-dimensional setting of the problem. Notably, in the opposite regime of very high-dimensional settings, the system variables may have universal statistical structures, such as well-defined eigenvalue spectra of their covariance matrices [30]. Looking forward, it would be intriguing to examine these universality properties in light of the perspectives developed in our work, seeking insights about the phase transition behavior for high-dimensional settings. In addition, it would be interesting to explore the possible implications of our results for settings beyond the canonical Gaussian information bottleneck. For instance, Ngampruetikorn and Schwab recently showed that similar structural transitions in the optimal representation occur when generalized correlation measures, such as Rényi and Jeffreys divergences, are used instead of the Shannon mutual information [28]. Whether the geometric and

componentwise information-theoretic criteria we derived for the transition points and the optimal navigation between them also hold for this generalized problem formulation remains an open question. It would also be fruitful to understand how and to what extent these perspectives on structural transitions continue to hold approximately when the variables are no longer jointly Gaussian [26,31,32], and what insights can be gained from the nature of the resulting deviations. When signal and relevance variables in such nonlinear systems represent entire trajectories rather than low-dimensional vectors, advanced simulation schemes may be required for the exact computation of information metrics [33].

ACKNOWLEDGMENTS

This work was supported by the Dutch Research Council (NWO) and the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Program (Grant No. 885065). It was performed at the research institute AMOLF.

DATA AVAILABILITY

The data that support the findings of this article are openly available [35].

[1] C. G. Bowsher and P. S. Swain, Environmental sensing, information transfer, and cellular decision-making, Curr. Opin. Biotechnol. **28**, 149 (2014).

[2] J. D. Victor, S. D. Boie, E. G. Connor, J. P. Crimaldi, G. B. Ermentrout, and K. I. Nagel, Olfactory navigation and the receptor nonlinearity, J. Neurosci. **39**, 3713 (2019).

[3] S. E. Palmer, O. Marre, M. J. Berry, and W. Bialek, Predictive information in a sensory population, Proc. Natl. Acad. Sci. USA **112**, 6908 (2015).

[4] E. Nelson, M. Corah, and N. Michael, Environment model adaptation for mobile robot exploration, Auton. Robots **42**, 257 (2018).

[5] A. Goyal, R. Islam, D. J. Strouse, Z. Ahmed, H. Larochelle, M. Botvinick, Y. Bengio, and S. Levine, Infobot: Transfer and exploration via the information bottleneck, in *International Conference on Learning Representations* (2019).

[6] N. Tishby, F. C. Pereira, and W. Bialek, The information bottleneck method, in *Proceedings of the 37th Allerton Conference on Communication and Computation* (University of Illinois, 1999).

[7] V. Sachdeva, T. Mora, A. M. Walczak, and S. E. Palmer, Optimal prediction with resource constraints using the information bottleneck, PLoS Comput. Biol. **17**, e1008743 (2021).

[8] A. J. Tjalma, V. Galstyan, J. Goedhart, L. Slim, N. B. Becker, and P. R. ten Wolde, Trade-offs between cost and information in cellular prediction, Proc. Natl. Acad. Sci. USA **120**, e2303078120 (2023).

[9] Z. Goldfeld and Y. Polyanskiy, The information bottleneck problem and its applications in machine learning, IEEE J. Sel. Areas Inf. Theory **1**, 19 (2020).

[10] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, Information bottleneck for Gaussian variables, J. Mach. Learn. Res. **6**, 165 (2005).

[11] W. H. de Ronde, F. Tostevin, and P. R. ten Wolde, Effect of feedback on the fidelity of information transmission of time-varying signals, Phys. Rev. E **82**, 031914 (2010).

[12] M. Chalk, O. Marre, and G. Tkačik, Toward a unified theory of efficient, predictive, and sparse coding, Proc. Natl. Acad. Sci. USA **115**, 186 (2018).

[13] R. E. Kalman, A new approach to linear filtering and prediction problems, J. Basic Eng. **82**, 35 (1960).

[14] A. Scaglione, P. Stoica, S. Barbarossa, G. B. Giannakis, and H. Sampath, Optimal designs for space-time linear precoders and decoders, IEEE Trans. Signal Process. **50**, 1051 (2002).

[15] H. Hotelling, Relations between two sets of variates, Biometrika **28**, 321 (1936).

[16] F. Creutzig, A. Globerson, and N. Tishby, Past-future information bottleneck in dynamical systems, Phys. Rev. E **79**, 041925 (2009).

[17] N. B. Becker, A. Mugler, and P. R. ten Wolde, Optimal prediction by cellular signaling networks, Phys. Rev. Lett. **115**, 258103 (2015).

[18] See Supplemental Material at http://link.aps.org/supplemental/10.1103/fnd1-8hty for details and additional studies, which includes Ref. [34].

[19] V. Ngampruetikorn and D. J. Schwab, Perturbation theory for the information bottleneck, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2021), Vol. 34, pp. 21008–21018.

[20] D. Mulder, P. R. ten Wolde, and T. E. Ouldridge, Exploiting bias in optimal finite-time copying protocols, Phys. Rev. Res. **7**, L012026 (2025).

[21] B. R. Gálvez, R. Thobaben, and M. Skoglund, The convex information bottleneck Lagrangian, Entropy **22**, 98 (2020).

[22] A. Kolchinsky, B. D. Tracey, and D. H. Wolpert, Nonlinear information bottleneck, Entropy **21**, 1181 (2019).

[23] R. M. Hecht, E. Noor, and N. Tishby, Speaker recognition by gaussian information bottleneck, in *Proceedings of Interspeech* (2009), pp. 1567–1570.

[24] A. Gordon, A. Banerjee, M. Koch-Janusz, and Z. Ringel, Relevance in the renormalization group and in information theory, Phys. Rev. Lett. **126**, 240601 (2021).

[25] A. G. Kline and S. E. Palmer, Gaussian information bottleneck and the non-perturbative renormalization group, New J. Phys. **24**, 033007 (2022).

[26] M. Rey and V. Roth, Meta-Gaussian information bottleneck, in *Advances in Neural Information Processing Systems* (2012), Vol. 25.

[27] M. M. Mahvari, M. Kobayashi, and A. Zaidi, Scalable vector Gaussian information bottleneck, in *IEEE International Symposium on Information Theory (ISIT)* (Melbourne, 2021).

[28] V. Ngampruetikorn and D. J. Schwab, Generalized information bottleneck for Gaussian variables, arXiv:2303.17762.

[29] S. Still, D. A. Sivak, A. J. Bell, and G. E. Crooks, Thermodynamics of prediction, Phys. Rev. Lett. **109**, 120604 (2012).

[30] V. Ngampruetikorn and D. J. Schwab, Information bottleneck theory of high-dimensional regression: Relevancy, efficiency and optimality, in *Advances in Neural Information Processing Systems* (2022), Vol. 35, pp. 9784–9796.

[31] M. Bauer and W. Bialek, Information bottleneck in molecular sensing, PRX Life **1**, 023005 (2023).

[32] T. Wu and I. Fischer, Phase transitions for the information bottleneck in representation learning, in *International Conference on Learning Representations* (2020).

[33] M. Reinhardt, G. Tkačik, and P. R. ten Wolde, Path weight sampling: Exact Monte Carlo computation of the mutual information between stochastic trajectories, Phys. Rev. X **13**, 041017 (2023).

[34] S. Särkkä and A. Solin, *Applied Stochastic Differential Equations*, Institute of Mathematical Statistics Textbooks (Cambridge University Press, 2019).

[35] V. Galstyan, Jupyter notebooks for reproducing the figures of the paper titled: "Intuitive dissection of the Gaussian information bottleneck method with an application to optimal prediction" (https://doi.org/10.1103/fnd1-8hty), Zenodo (2025), available at: https://doi.Org/10.5281/zenodo.17495525.