SYNOPTIC INFORMATION: A VARIATIONAL PRINCIPLE FOR JOINT REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

How should two high-dimensional noisy views of the same system be jointly compressed so that their representations capture structure common to both while flexibly discarding view-specific detail? Learning such a joint reduced representation by exploiting shared internal structure obviates the need for an external teacher to extract interesting features from the input data. Here, an information-theoretic description of the problem is introduced and formalized as a two-temperature variational principle for the synoptic information, the optimal number of bits shared between the two compressed views that maximally retain what each representation predicts about the other. Functional optimization yields exact coupled self-consistent equations for both encoders, readily solved by a generalized iterative algorithm. For Gaussian distribution inputs, the synoptic information decomposes as a sum over canonical correlation modes, admitting closed-form analytical solutions. The phase diagram of optimal solutions in the temperature plane reveals phase transitions through nested regions of active modes and continuously interpolates between differing representation learning frameworks, such as canonical correlation analyses and information bottleneck, which appear as limiting cases.

1 INTRODUCTION

A common theme in modern representation learning is to encode two views of the same data whilst rewarding the agreement between the two encoders. The two views could be differing modalities, or the same modality at differing points in space and time. Recent frameworks such as contrastive learning (Chen et al., 2020), CLIP (Radford et al., 2021), Barlow Twins (Zbontar et al., 2021), VICReg (Bardes et al., 2022) and joint-embedding predictive architectures (LeCun, 2022), all follow this approach, albeit with differing heuristics. However, none of them explicitly state how much each view should be compressed or why the answer should differ between a rich view and a noisy one.

From an information-theoretic perspective, the principle of maximizing the shared information between paired observations originates with Imax (Becker & Hinton, 1992), which proposed training neural network modules to maximize the mutual information between adjacent image patches. This established the concept that shared structure in coupled data can supervise representation learning without external labels. With Imax the compression is not explicitly modeled but is fixed by the architecture. The information bottleneck (Tishby et al., 1999) provided a principled framework to explicitly model one-sided compression while preserving information about a fixed target with a temperature parameter controlling the tradeoff.

The present work introduces a *two-temperature* variational formulation that independently controls the compression on each side. We define the quantity *synoptic information* [1] to describe the shared optimal information in the joint representation space of the two views. Introducing independent temperatures is necessary to span the full landscape of compressions where both sources may carry information at substantially different rates, a generic situation in multi-modal learning where one modality is intrinsically richer or noisier than the other. For example, in functional genomics, a common goal is to understand what features of the DNA sequences are associated with function, and vice versa. Functions such as gene expression exhibit greater stochasticity and sparsity than the

[1] The word synoptic means seeing multiple things from a single point of view, and is derived from the Greek *synoptikos* which combines syn- (together) and opsis (seeing or view)

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

concomitant DNA sequencing data, and thus requiring a greater amount of compression to seek a useful joint latent space.

Contributions. (i) We introduce the synoptic information formalism with two independent temperatures θ_x and θ_y and derive coupled self-consistent equations for the optimal encoders (Section 2). (ii) For multivariate Gaussian input, exact solutions are obtained and analyzed (Section 3). (iii) The rich structure of optimal solutions is presented, revealing structural phase transitions in the reduced representation and how differing representation learning paradigms emerge from the synoptic information framework. (Section 3.5).

2 THE SYNOPTIC INFORMATION FRAMEWORK

2.1 SETUP AND OBJECTIVE

Let X and Y be random variables with joint distribution $p(x, y)$, representing two views of a common system. We seek stochastic encoders $p(s_x|x)$ and $p(s_y|y)$ that map each view into a compressed representation, S_x and S_y, such that the representations are maximally informative about one another while individually retaining as little as possible about the raw input. Neither encoder has direct access to the other. The Markov structure for this formalism is

$$S_x \leftrightarrow X \leftrightarrow Y \leftrightarrow S_y. \tag{1}$$

Under this condition, the joint distribution of the variables factorizes as $p(x, y, s_x, s_y) = p(s_x|x)p(s_y|y)p(x, y)$. To quantify the shared information between variables, say X and Y, we use mutual information $I(X; Y)$, a fundamental concept in information theory (Shannon, 1948) and a self-equitable measure of association (Kinney & Atwal, 2014). From the Data Processing Inequality applied to the above Markov condition we note that $I(S_x; S_y) \leq I(X; Y)$, capturing our intuition that the shared information in the compressed representation cannot be greater than between the original sources with equality when there is no compression.

We formalize our objective of finding the optimal encoding by maximizing the following functional,

$$\mathcal{F} \;=\; I(S_x;\, S_y) \;-\; \theta_x\, I(S_x;\, X) \;-\; \theta_y\, I(S_y;\, Y)\,, \tag{2}$$

over the pair of encoders $p(s_x|x)$ and $p(s_y|y)$, with $\theta_x, \theta_y > 0$. The first term rewards shared information between the two compressed representations. The second and third terms penalize the information that each representation retains about its own input, weighted by independent temperatures, θ_x and θ_y, serving as Lagrange parameters in the constrained optimization problem. The higher the temperature, the greater the compression from the input space to the compressed representation, or latent, space. Note the negative functional has the form of the free energy in statistical physics, specifying the temperature-weighted tradeoff between energy (shared information) and entropy (compressive information) terms. The equilibrium distribution of states is determined by minimizing the free energy, corresponding to maximizing the functional in Eq.(2).

The *synoptic information* (SI) is defined as the optimal shared information between the two representations, $I(S_x^*; S_y^*)$, where the optimal representations (S_x^*, S_y^*) are determined by maximizing the functional

$$(S_x^*, S_y^*) \;=\; \underset{p(s_x|x),\, p(s_y|y)}{\arg\max} \Big\{ I(S_x; S_y) - \theta_x\, I(S_x; X) - \theta_y\, I(S_y; Y) \Big\}. \tag{3}$$

When $\theta_y = 0$ the Y-encoder is unconstrained, $S_y = Y$, and the objective reduces to $I(S_x;\, Y) - \theta_x\, I(S_x;\, X)$, which is equivalent to the standard Information Bottleneck (IB) functional (Tishby et al., 1999). The original X variable is compressed into a reduced representation S_x while retaining as much information about the relevant variable Y. A similar argument applies in the $\theta_x = 0$ limit, resulting in an IB functional for the compressed Y variable, $I(S_y;\, X) - \theta_y\, I(S_y;\, Y)$. Thus the synoptic information formalism provides a single, continuously parameterized objective that interpolates between two IB frameworks.

2.2 SELF-CONSISTENT EQUATIONS

Setting the functional derivative $\delta\mathcal{F}/\delta p(s_x|x) = 0$ subject to the normalization constraint $\sum_{s_x} p(s_x|x) = 1$ gives the stationarity condition for the X-encoder. The derivation parallels that of

the standard IB (Tishby et al., 1999), but with the additional challenge that the variation of $p(s_x|x)$ affects $I(S_x;\, S_y)$ through two paths, directly through $p(s_x)$ and indirectly through $p(s_y|s_x)$. The indirect variation vanishes identically by normalization of $p(s_y|s_x)$, for the same reason it vanishes in the IB (Tishby et al., 1999). The result is:

$$p(s_x|x) \;=\; \frac{p(s_x)}{Z(x,\theta_x)}\,\exp\!\left(-\frac{1}{\theta_x}D_{KL}[p(s_y|x)||p(s_y|s_x)]\right), \tag{4}$$

$$p(s_y|y) \;=\; \frac{p(s_y)}{Z(y,\theta_y)}\,\exp\!\left(-\frac{1}{\theta_y}D_{KL}[p(s_x|y)||p(s_x|s_y)]\right), \tag{5}$$

where $D_{KL}[p||q]$ is the Kullback-Leibler divergence between the two distributions, and Z is the normalization constant or partition function. Note that this solution for the optimal encoder has the form of a Boltzmann distribution where the Kullback-Leibler divergence emerges as the effective energy or distortion. The probabilistic assignment of s_x given x is determined by the compressed Y representation through how closely $p(s_y|x)$ matches $p(s_y|s_x)$. The temperature factor θ_x controls the compression rate and consequently the sharpness of the encoder with a smaller θ_x producing a more deterministic mapping.

The two equations Eqs. 4–5 are coupled: the X-encoder depends on the conditional distribution $p(s_y|s_x)$, which is determined by the Y-encoder, and vice versa. They also involve the marginals $p(s_x)$ and $p(s_y)$, which are themselves functions of the encoders. This structure admits a natural alternating optimization scheme.

2.3 Generalized Blahut–Arimoto algorithm

We solve Eqs. 4–5 by alternation: given the current Y-encoder, update $p(s_x|x)$ according to 4, recompute the joint statistics, update $p(s_y|y)$ according to 5, iterate. Each half-step can be shown to increase $\mathcal{F}$ by a standard argument extending the Blahut–Arimoto convergence proof (Blahut, 1972; Arimoto, 1972) to the two-encoder setting, since each update maximizes $\mathcal{F}$ with respect to one encoder while holding the other fixed.

A practical difficulty is that the trivial solution $p(s_x|x) = p(s_x)$, $p(s_y|y) = p(s_y)$, in which the representations carry no information, is always a fixed point. We avoid this by deterministic annealing: begin at small θ_x, θ_y (where the $1/\beta$ factors in the exponents make the nontrivial fixed point strongly attracting) and gradually increase both temperatures to their target values. This traces a continuous path through the space of solutions, analogous to the annealing procedure used in the standard IB (Tishby et al., 1999).

3 Gaussian analytical solution

The information theoretic nature of the SI formulation lends itself to deal with both categorical and continuous data. Here we address the problem of modeling Gaussian processes, an important generative class in machine learning. When (X, Y) are jointly Gaussian, the SI formalism admits a complete closed-form solution, revealing a detailed understanding of continuous dimensionality reduction and structure of solutions across the plane of both temperatures. The theory generalizes the Gaussian information bottleneck of Chechik et al. (2005) to two encoders and reveals a richer phase structure. We find that activation occurs along curves, bifurcations are first-order in the interior of the temperature plane and second-order only on the IB-axis edges.

3.1 Setup and Outline

Let $X \in \mathbb{R}^{d_x}$ and $Y \in \mathbb{R}^{d_y}$ be jointly Gaussian with zero mean and respective dimensions d_X and d_Y. Input data across dimensions are coupled, measured by covariances $\mathrm{Cov}(X) = \Sigma_X \in \mathbb{R}^{d_X \times d_X}$ and $\mathrm{Cov}(Y) = \Sigma_Y \in \mathbb{R}^{d_Y \times d_Y}$, and cross covariance $\mathrm{Cov}(X,Y) = \Sigma_{XY} \in \mathbb{R}^{d_X \times d_Y}$. The encoders are linear maps with additive Gaussian noise,

$$S_X = AX + \xi_X, \qquad S_Y = BY + \xi_Y, \qquad \xi_X \sim \mathcal{N}(0, \Sigma_{\xi_X}),\;\; \xi_Y \sim \mathcal{N}(0, \Sigma_{\xi_Y}), \tag{6}$$

where $A \in \mathbb{R}^{k_X \times d_X}$ and $B \in \mathbb{R}^{k_Y \times d_Y}$ are the encoder matrices, k_X and k_Y are the dimensions of the two representations, and the noises are independent of each other and of (X, Y).

The goal of finding the optimal encoders is equivalent to finding the transformation matrices A and B that maximize the SI functional Eq.(2). This can be a challenging numerical task, but it will be shown that an exact analytical solution can be determined for Gaussian views by proceeding in three steps. i) *Choose independent coordinates*. Mutual information is invariant to invertible transformations, so there is freedom to rotate and rescale the joint Gaussian (X, Y) into canonical coordinates where the two correlated views are decomposed into a set of independent pairs of scalars $(\tilde{X}_i, \tilde{Y}_i)$. ii) *Align encoders along the independent coordinates*. This procedure reduces the multi-dimensional encoding problems into a set of simpler one-dimensional tasks. Instead of sending all X directly through a noisy channel to the representation S_X, each canonical variable $\tilde{X}_i$ is sent separately through its own noisy channel to $(S_X)_i$. The same applies to the Y variable. iii) *Optimize the aligned encoders, one mode at a time*. Since the SI functional becomes a sum over independent modes, each noisy channel can be separately optimized.

3.2 Canonical coordinates

Mutual information between joint variables is invariant to invertible transformations of either variable, $I(X;Y) = I(f(X);g(Y))$, so every term of the SI functional, Eq.(2), is unchanged if we replace X, Y, S_X and S_Y by their transformed counterparts. We can use this freedom to choose new coordinates of the joint distribution $p(X, Y)$ that simplifies the subsequent SI optimization. The goal will be to decompose the full complex SI functional into a simpler set of one-dimensional functionals by finding joint coordinates that are independent of each other.

For Gaussian variables, finding independent features is equivalent to decorrelating the cross-covariance. To this end, we first whiten each view, $\tilde{X} = \Sigma_X^{-1/2} X$ and $\tilde{Y} = \Sigma_Y^{-1/2} Y$, to give identity covariances. Whitening fixes the within-view structure by rescaling the individual variances and making each view look like decorrelated noise on its own. Their cross-covariance is the matrix

$$C = \Sigma_X^{-1/2}\, \Sigma_{XY}\, \Sigma_Y^{-1/2}, \tag{7}$$

which, in general, is a full matrix whereby $\tilde{X}$ and $\tilde{Y}$ can still be correlated. To fix the between-view structure we carry out a singular value decomposition $C = U\mathsf{R}V^\top$, with U, V orthogonal and $\mathsf{R} \in \mathbb{R}^{d_X \times d_Y}$ zero everywhere except for $\mathsf{R}_{ii} = \rho_i$, $i = 1, \ldots, r$, where $1 > \rho_1 \geq \cdots \geq \rho_r > 0$ and $r = \operatorname{rank} \Sigma_{XY}$. Finally we rotate, $\hat{X} = U^\top \tilde{X}$ and $\hat{Y} = V^\top \tilde{Y}$, into canonical variables, and in these coordinates

$$\mathrm{Cov}(\hat{X}) = I_{d_X}, \qquad \mathrm{Cov}(\hat{Y}) = I_{d_Y}, \qquad \mathrm{Cov}(\hat{X}, \hat{Y}) = \mathsf{R}. \tag{8}$$

We arrive at the goal of decorrelating both the within-view and between-view structure. $\hat{X}_i$ is correlated only with $\hat{Y}_i$, with correlation coefficient ρ_i, and is independent of every other coordinate of both views. This process underlies canonical correlation analysis (CCA) whereby linear combinations of X and Y are sought that have maximal correlation with each other (Hotelling, 1936).

The joint Gaussian has been decomposed into r independent *canonical modes*, pairs of unit-variance scalars $(\hat{X}_i, \hat{Y}_i)$ with correlation ρ_i, plus leftover coordinates carrying no information about the other view. For scalar Gaussian variables with correlation ρ the mutual information has a simple expression, $I = -\frac{1}{2}\log(1-\rho^2)$ (Shannon, 1948). From the independence of modes it follows that the mutual information of multivariate Gaussian variables decomposes as a sum over canonical modes,

$$I(X;Y) = -\tfrac{1}{2} \sum_{i \leq r} \log(1 - \rho_i^2) \tag{9}$$

Consequently, we will show that the synoptic functional can similarly be represented a sum of independent scalar functions. Finally, replacing S_X by $\Sigma_{\xi_X}^{-1/2} S_X$ changes nothing, so we take unit noise, $\xi_X \sim \mathcal{N}(0, I_{k_X})$ and $\xi_Y \sim \mathcal{N}(0, I_{k_Y})$, and from here on let A and B act on the canonical variates: $S_X = A\hat{X} + \xi_X$, $S_Y = B\hat{Y} + \xi_Y$.

3.3 Aligned encoders

In canonical coordinates, we let the encoders treat each canonical variate separately by sending $\hat{X}_i$ through its own noisy channel with some gain, and ignore the leftover coordinates,

$$(S_X)_i = \sqrt{u_i}\, \hat{X}_i + \xi_i, \qquad (S_Y)_i = \sqrt{v_i}\, \hat{Y}_i + \eta_i, \qquad \xi_i, \eta_i \sim \mathcal{N}(0,1) \text{ independent.} \tag{10}$$

Since $\hat{X}_i$ has unit variance and the noise has unit variance, $u_i \geq 0$ is the signal-to-noise ratio (SNR) at which the i-th canonical variate of X is transmitted, and v_i plays the same role for Y. The covariance matrix of S_X with S_Y takes on a simple form that depends only on the correlation and SNRs,

$$\Sigma_{(S_X)_i,(S_Y)_i} = \begin{pmatrix} 1+u_i & \rho_i\sqrt{u_i v_i} \\ \rho_i\sqrt{u_i v_i} & 1+v_i \end{pmatrix} \tag{11}$$

In the original coordinates the encoders are expressed as

$$A = \operatorname{diag}(\sqrt{u_1}, \sqrt{u_2}, \dots)\, U^\top \Sigma_X^{-1/2}, \qquad B = \operatorname{diag}(\sqrt{v_1}, \sqrt{v_2}, \dots)\, V^\top \Sigma_Y^{-1/2}: \tag{12}$$

The rows of A are the CCA projection vectors of X, each scaled by its gain. Everything about the encoders is thus captured by the two lists of SNRs.

To decompose the SI functional into canonical modes we note that each of the three terms of the SI functional, Eq.2, is a sum over i, and each summand is a mutual information between scalar Gaussians. Within mode i: the correlation between $(S_X)_i$ and $\hat{X}_i$ is $\sqrt{u_i/(1+u_i)}$, so $I\big((S_X)_i; \hat{X}_i\big) = \frac{1}{2}\log(1+u_i)$; likewise $I\big((S_Y)_i; \hat{Y}_i\big) = \frac{1}{2}\log(1+v_i)$; and the correlation between $(S_X)_i$ and $(S_Y)_i$ is

$$r_i = \rho_i \sqrt{\frac{u_i}{1+u_i}} \sqrt{\frac{v_i}{1+v_i}}, \tag{13}$$

the canonical correlation attenuated by the two channels. Altogether,

$$\mathcal{F} = \sum_{i=1}^{r} f_i(u_i, v_i), \qquad f_i(u_i, v_i) = -\frac{1}{2}\log\Big(1 - \rho_i^2 \frac{u_i}{1+u}\frac{v}{1+v}\Big) - \frac{\theta_x}{2}\log(1+u) - \frac{\theta_y}{2}\log(1+v). \tag{14}$$

The two-encoder multivariate problem has become r independent scalar problems.

Suppose the Y-encoder is fixed. The X-side problem is a one-dimensional Gaussian IB with S_Y as relevance variable, and the correlation between $\hat{X}$ and S_Y is $\rho\sqrt{v/(1+v)}$. The Gaussian IB has a simple threshold (Chechik et al., 2005): a direction is worth encoding only if the temperature is below its squared correlation with the relevance variable. Applied to both sides,

$$X \text{ encodes the mode only if } \theta_x < \rho^2 \frac{v}{1+v}, \qquad Y \text{ encodes the mode only if } \theta_y < \rho^2 \frac{u}{1+u}. \tag{15}$$

The ability of each side to resolve a mode is determine by how well the other side has already resolved the mode. Near the trivial solution $u = v = 0$ neither inequality can hold, however small the temperatures: the encoders must reach a finite resolution together before either finds the mode worth its cost. This has no counterpart in the IB, where the relevance variable is given rather than learned.

3.4 One mode in closed form

Since the modes do not interact it suffices to solve for a single mode with an arbitrary canonical correlation ρ. The full solution is then a sum over modes. It is convenient to write the SI functional in terms of the determinant of covariance matrix of (S_X, S_Y) Eq.11

$$D \equiv \det \Sigma_{S_X,S_Y} = (1+u)(1+v) - \rho^2 uv, \tag{16}$$

so

$$f(u,v) = \tfrac{1}{2}\log\frac{(1+u)(1+v)}{D} - \frac{\theta_x}{2}\log(1+u) - \frac{\theta_y}{2}\log(1+v). \tag{17}$$

Upon optimizing the SI functional, by setting $\partial f/\partial u = \partial f/\partial v = 0$, we find solutions for the optimal SNRs and criteria for when the mode is active. In this section we report the main results, and leave the detailed derivations to the Appendix A.

1. At stationarity, optimal ratio of SNRs is determined by the ratio of temperatures,

$$\frac{u}{v} = \frac{\theta_y}{\theta_x}. \tag{18}$$

The side that is cheaper to compress transmits at the higher SNR.

2. A non-trivial stationary point exists when

$$\rho^2 - \theta_x - \theta_y \;\geq\; 2\sqrt{(1-\rho^2)\,\theta_x\theta_y}\,. \tag{19}$$

3. The optimal D has two solutions, and the corresponding SNRs, are

$$\begin{aligned} D_\pm &= \frac{\rho^2\Big[(\rho^2-\theta_x-\theta_y) \pm \sqrt{(\rho^2-\theta_x-\theta_y)^2 - 4(1-\rho^2)\theta_x\theta_y}\Big]}{2(1-\rho^2)\theta_x\theta_y}, \\ u_\pm &= \frac{\theta_y D_\pm}{\rho^2}, \qquad v_\pm = \frac{\theta_x D_\pm}{\rho^2}; \end{aligned} \tag{20}$$

D_+ is a local maximum of f and D_- a saddle. The mode is active, i.e. the global maximum is at D_+, if and only if Eq.19 holds and

$$f_+ \;:=\; \tfrac{1}{2}\Big[(1-\theta_x)\log(1+u_+) + (1-\theta_y)\log(1+v_+) - \log D_+\Big] \;>\; 0. \tag{21}$$

Otherwise the optimum is the trivial encoder $u = v = 0$.

Condition Eq.19 says when a non-trivial encoder exists as a local optimum; condition Eq.21 says when it beats the trivial encoder. In the language of first-order phase transitions, Eq.19 is the *spinodal* and Eq.21 the *coexistence* condition; the region defined by Eq.21 lies strictly inside the one defined by Eq.19, and in between the non-trivial encoder is metastable.

Since the trivial encoder is always a local maximum, the optimum cannot move continuously away from it. Where the two roots of Eq.20 merge, the non-trivial encoder is born at a finite SNR,

$$u^* v^* = \frac{1}{1-\rho^2}, \qquad \frac{u^*}{v^*} = \frac{\theta_y}{\theta_x}, \tag{22}$$

and on the coexistence boundary the jump is larger still. On the line $\theta_x = \theta_y$ a mode appears with $u^* = v^* = (1-\rho^2)^{-1/2} \geq 1$.

As $\theta_y \to 0$ the Y-encoder becomes lossless ($v_+ \to \infty$), Eq. 19 reduces to $\theta_x \leq \rho^2$, the two thresholds coalesce and the jump vanishes. On the axis the mode switches on continuously at $\theta_x = \rho^2$, the Gaussian IB critical temperature (Section 3.6). Numerically the jump closes as $\sqrt{\theta_y}$ up to slowly varying corrections. The continuous IB transition is thus the endpoint of a line of discontinuous ones, as a critical point terminates a liquid–gas coexistence line.

Writing $\mathcal{A}$ for the set of active modes, the optimal encoders achieve $\mathcal{F}^\star = \sum_{i\in\mathcal{A}} f_i^+$, $I(S_X;X) = \frac{1}{2}\sum_{i\in\mathcal{A}}\log(1+u_i^+)$ and $I(S_Y;Y) = \frac{1}{2}\sum_{i\in\mathcal{A}}\log(1+v_i^+)$, with u_i^+, v_i^+ from Eq.20 at $\rho = \rho_i$.

3.5 The phase diagram

For each canonical mode i define its *active region*

$$\Omega_i := \big\{(\theta_x,\theta_y) \in (0,1)^2 :\; f_i^+(\theta_x,\theta_y) > 0\big\}. \tag{23}$$

The regions are nested, $\Omega_1 \supseteq \Omega_2 \supseteq \cdots \supseteq \Omega_r$, because a mode with larger canonical correlation is always resolved at least as finely. Swapping the optimal SNRs of two modes cannot improve either, the compression costs cancel under the swap, and the shared term rewards resolution more strongly for the larger ρ; with $u/v = \theta_y/\theta_x$ fixed by Eq. 18 this gives $u' \geq u$ and $v' \geq v$, so whenever mode i is active every mode $j < i$ is active too. Thus the number of active modes

$$k(\theta_x,\theta_y) = \#\{i : (\theta_x,\theta_y) \in \Omega_i\}$$

is a staircase descending from r near the origin to 0 outside Ω_1 (Figure 1 and Figure 2). On each step the optimal representation is the aligned encoder Eq.12 on the first k modes, with SNRs given by Eq.18–Eq.20. Crossing a coexistence curve adds one mode, discontinuously and every Ω_i meets the axes at $\theta = \rho_i^2$.

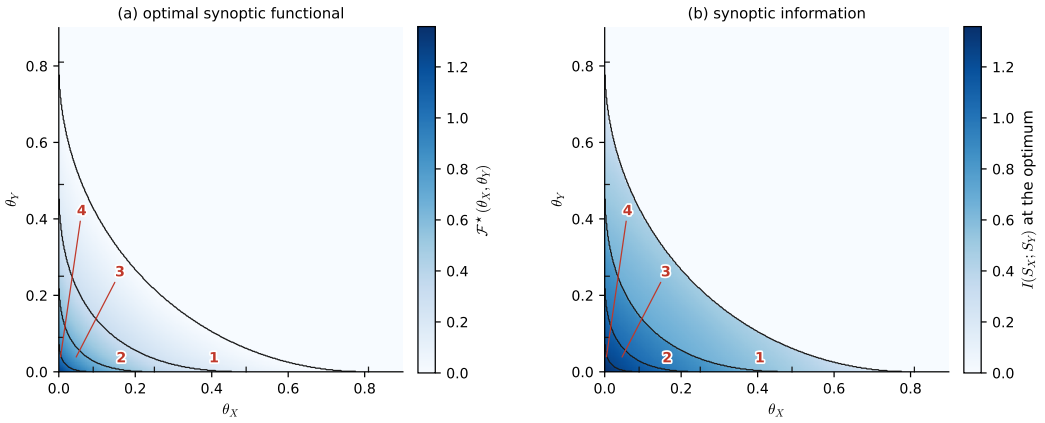


Figure 1: **Synoptic information in the temperature plane** for Gaussian views with canonical correlations $\rho = (0.9, 0.7, 0.5, 0.3)$. (a) The optimal synoptic functional $\mathcal{F}^\star(\theta_x, \theta_y)$. (b) The synoptic information $(S_X; S_Y)$ of the optimal encoders. Black curves: the boundaries $f_i^+ = 0$ at which mode i switches on. The numbers give the number of active modes in each region. The regions are nested and each meets the axes at $\theta = \rho_i^2$ (ticks), the Gaussian IB critical temperature. $\mathcal{F}^\star$ is continuous across every boundary. $(S_X; S_Y)$ jumps across every boundary in the interior and is continuous only on the axes. Both are symmetric about the diagonal.

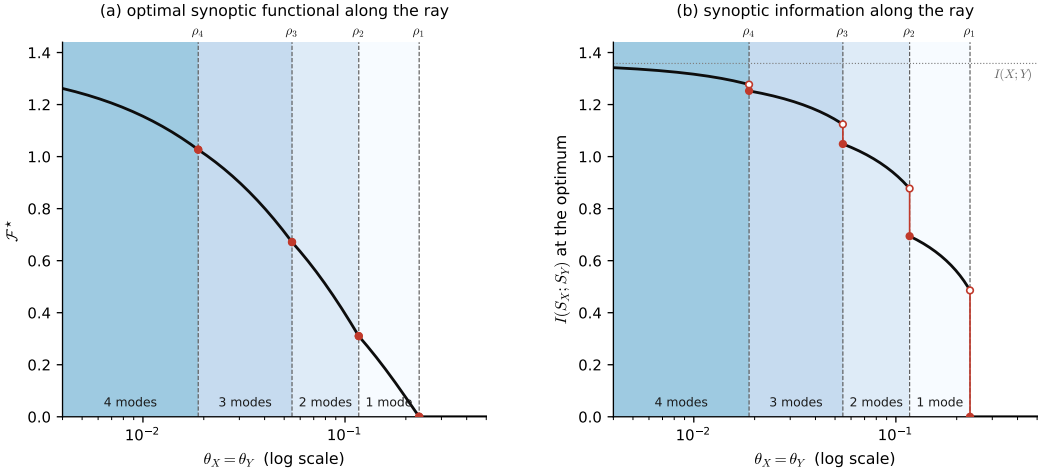


Figure 2: **Mode transitions along the symmetric ray** $\theta_x = \theta_y$ for the source of Figure 1 (log temperature axis). Shading: number of active modes. Dashed lines: the onsets of modes 4, 3, 2, 1 (labelled by their canonical correlations $\rho_4 < \rho_3 < \rho_2 < \rho_1$) as the temperature is lowered from right to left. (a) The optimal functional $\mathcal{F}^\star$ is continuous through every onset (red points) but its slope changes there: by the envelope relation $\partial\mathcal{F}^\star/\partial\theta_x = -I(S_X; X)$, the kinks in (a) are the jumps of the compression information. (b) The synoptic information $I(S_X; S_Y)$ jumps at every onset (red: the two sides of each jump). The dotted line is $I(X; Y)$, approached as $\theta \to 0$.

3.6 Limiting cases and relations to other frameworks

The temperature plane interpolates between several representation-learning principles, which emerge on its edges and corners.

Gaussian information bottleneck (the axes). On $\theta_y = 0$ the Y-encoder is lossless and equation 2 becomes $(S_X; Y) - \theta_x(S_X; X)$, the IB Lagrangian with $\theta_x = 1/\beta_{\mathrm{IB}}$, where β_{IB} is the trade-off parameter of Tishby et al. (1999). Mode i is active iff $\theta_x < \rho_i^2$ and the solution is that of Chechik

et al. (2005), whose eigenvalues λ_i of $\Sigma_{X|Y}\Sigma_X^{-1}$ equal $1-\rho_i^2$. The axis $\theta_x = 0$ is the mirror image. The axes are the "predictive" and "generative" extremes in which only one view is compressed.

CCA and Imax (the origin). As $\theta_x, \theta_y \to 0$ all r modes are active with $u_i, v_i \to \infty$, the noise becomes negligible, Eq.12 becomes the deterministic CCA transformation and $\mathcal{F} \to I(X;Y)$. This is the linear-Gaussian instance of the Imax principle of Becker & Hinton (1992), which maximizes the mutual information between the outputs of two modules seeing different parts of the input and reduces to CCA for linear Gaussian modules (Becker, 1996). Synoptic information is a two-temperature deformation of Imax that penalizes the information each module keeps about its own input. The temperatures regularize CCA by discarding weakly correlated modes and transmitting the rest at finite precision, and the aligned encoder Eq.12 is the noisy-channel form of probabilistic CCA (Bach & Jordan, 2005).

Symmetric compression (the diagonal). On $\theta_x = \theta_y = \theta$ the two SNRs are equal and the problem is a Gaussian instance of the symmetric information bottleneck (Slonim et al., 2006) and the double-sided information bottleneck (Dikshtein et al., 2022). Condition Eq. 19 becomes $4\theta(1-\theta) \le \rho^2$, i.e. $\theta \le \frac{1}{2}(1-\sqrt{1-\rho^2}) \approx \rho^2/4$ for small ρ: with both sides paying, a mode survives only to about a quarter of the IB critical temperature.

Self-supervised and multi-view learning. The same corners organize several recent objectives: contrastive methods (Chen et al., 2020), Barlow Twins and VICReg (Zbontar et al., 2021; Bardes et al., 2022) sit at the Imax corner (the latter recover CCA-type solutions in their linear regime, Balestriero & LeCun, 2022), the multi-view IB of Federici et al. (2020) penalizes view-specific information through conditional terms such as $I(S_X; X|Y)$, and joint-embedding predictive architectures (LeCun, 2022) learn two encoders and a predictor with anti-collapse regularizers in the role of the temperatures.

4 Discussion

The synoptic information methodology provides a pair of representations that are informative about each other and about nothing else. The two temperatures are not redundant since in the Gaussian solution the ratio θ_y/θ_x fixes the relative fidelity $u_i/v_i = \theta_y/\theta_x$ on every active mode, while the distance from the origin along a ray of fixed ratio sets how many modes survive. Moving along a ray therefore changes what is represented (the number of canonical modes). Moving across rays changes how the representational burden is split between the two sides. For non-Gaussian $p(x,y)$ this decoupling into a ratio and a distance is lost, but the qualitative picture remains. Each side refines on its own schedule, and the regions of the temperature plane in which S_x and S_y change resolution need not coincide. A single-temperature formulation, such as the symmetric IB (Slonim et al., 2006) can only explore the diagonal of this plane.

At every temperature and for every $p(x,y)$ the trivial encoder (constant S_x, constant S_y) is a strict local maximum of $\mathcal{F}$: the compression costs are linear in a small perturbation of the encoders, while the reward $I(S_x;S_y)$ is bilinear and therefore of higher order. Nontrivial representations consequently appear only through first-order transitions, at which a finite representation abruptly becomes preferable to none, and every self-consistent solution of the coupled encoder equations must be reached from an initial condition that already sits on the far side of the barrier. This is, we believe, the information-theoretic content of "collapse" in self-supervised learning. The schemes that prevent it in practice (e.g. negative samples, stop-gradients, predictor asymmetry, variance and covariance regularizers) are different ways of placing the optimizer beyond the trivial basin. The annealed generalized Blahut-Arimoto iteration of Section 2 makes the barrier visible rather than hiding it. On the IB axes the barrier vanishes, the transition becomes continuous at $\theta = \rho^2$, and the familiar IB critical point is recovered as the endpoint of a line of first-order transitions.

Since every step of the Gaussian analysis is explicit (whitening, canonical coordinates, a closed-form single-mode functional, and a per-mode critical temperature) it serves as a useful testing ground in which the general theory can be checked. The phase diagram it predicts is a set of nested regions ordered by the canonical correlations, symmetric about the diagonal. Familiar methods sit on its boundaries: IB on the axes, symmetric and double-sided IB on the diagonal, Imax and CCA at the origin. The Gaussian case is symmetric around the diagonal ray on the temperature plane.

Asymmetry appears in non-Gaussian data, where the maximal-correlation spectrum takes over the role of the canonical correlations in governing the order in which structure is admitted.

The optimality of Gaussian channels among all encoders of a jointly Gaussian source is established here only within the linear-Gaussian class. The extension to arbitrary stochastic encoders, known for IB, remains open for the two-sided functional. The discrete solver requires multiple structured initializations and warm starts along the annealing path, since random initialization routinely misses inner phases, and it operates on tabulated joint distributions rather than raw high-dimensional inputs. A neural or variational implementation of $\mathcal{F}$ is the natural next step. The phase structure derived here provides a concrete set of predictions, such as first-order onsets and nested activation, that can be tested against future implemementations.

5 Conclusion

The synoptic information formalism has been introduced as a variational principle for learning two representations jointly by maximizing their mutual information while constraining each side, at its own temperature, for what it retains of its input. The principle admits a self-consistent characterization of its optima and an annealed iterative solver, and it contains the information bottleneck, its symmetric and double-sided variants, Imax, and CCA as boundary cases of one temperature plane. In the Gaussian setting it is exactly solvable. The optimal encoders align with the canonical modes, each mode switches on at a computable temperature through a first-order transition, and the resulting phase diagram is a nested family of regions in which the number of retained modes is set by the distance along a ray and their relative fidelity by its slope. The same analysis shows that the collapsed representation is always locally optimal, so that representation learning under this principle is a matter of crossing phase transitions. Thus, it is important to survey the entire temperature plane, rather than any single point in it, when asking what a pair of modalities have to say about each other.

References

Suguru Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18(1):14–20, 1972.

Francis R Bach and Michael I Jordan. A probabilistic interpretation of canonical correlation analysis. 2005.

Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. *Advances in Neural Information Processing Systems*, 35:26671–26685, 2022.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Suzanna Becker. Mutual information maximization: models of cortical self-organization. *Network: Computation in neural systems*, 7(1):7, 1996.

Suzanna Becker and Geoffrey E Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163, 1992.

Richard E. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, 18(4):460–473, 1972.

Gal Chechik, Amir Globerson, Naftali Tishby, and Yair Weiss. Information bottleneck for Gaussian variables. *Journal of Machine Learning Research*, 6:165–188, 2005.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Michael Dikshtein, Or Ordentlich, and Shlomo Shamai. The double-sided information bottleneck function. *Entropy*, 24(9):1321, 2022. doi: 10.3390/e24091321.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. *arXiv preprint arXiv:2002.07017*, 2020.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

Justin B. Kinney and Gurinder S. Atwal. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9):3354–3359, 2014. doi: 10.1073/pnas.1309933111.

Yann LeCun. A path towards autonomous machine intelligence. *OpenReview preprint*, 2022. Version 0.9.2, 2022-06-27.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Noam Slonim, Nir Friedman, and Naftali Tishby. Multivariate information bottleneck. *Neural computation*, 18(8):1739–1789, 2006.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, pp. 368–377, 1999.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 2021.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic data sets and assist in the writing of proofs. We have not used generative AI tools for developing theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, or interpreting results, and the rest of the required disclosure tasks are not applicable to this work.

Additionally, we used generative AI tools for create or modify scientific figures or images, create or edit software code, identify relevant literature. We have reviewed all AI-assisted work. LLM-generated code was verified for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

A THE SINGLE-MODE CALCULATION

Throughout, $D = (1+u)(1+v) - \rho^2 uv$ as in equation 16. Two partial derivatives of D will be used repeatedly:

$$D_u := \frac{\partial D}{\partial u} = 1 + (1-\rho^2)v, \qquad D_v := \frac{\partial D}{\partial v} = 1 + (1-\rho^2)u. \tag{24}$$

(i) Stationary points. Differentiating equation 17 with respect to u,

$$\frac{}{\partial u} = \frac{1}{2}\Big[\frac{1-\theta_x}{1+u} - \frac{D_u}{D}\Big]. \tag{25}$$

Setting this to zero and multiplying out gives $(1-\theta_x)\,D = (1+u)\,D_u$. Expanding the right-hand side,

$$(1+u)D_u = (1+u)(1+v) - \rho^2 v(1+u) = D + \rho^2 uv - \rho^2 v - \rho^2 uv = D - \rho^2 v, \tag{26}$$

so the stationarity condition simplifies to $\rho^2 v = \theta_x D$. By symmetry, $/\partial v = 0$ gives $\rho^2 u = \theta_y D$. Dividing the two, $u/v = \theta_y/\theta_x$. To find D itself, substitute $u = \theta_y D/\rho^2$ and $v = \theta_x D/\rho^2$ into the definition $D = 1 + u + v + (1-\rho^2)uv$:

$$D = 1 + \frac{(\theta_x + \theta_y)D}{\rho^2} + \frac{(1-\rho^2)\theta_x\theta_y D^2}{\rho^4}. \tag{27}$$

(ii) Existence of the roots, and which root is the maximum. The quadratic for D has positive leading coefficient and positive constant term, so its two roots have positive product. They are real and positive exactly when the discriminant is non-negative and their sum is positive:

$$(\rho^2 - \theta_x - \theta_y)^2 \geq 4(1-\rho^2)\theta_x\theta_y \qquad \text{and} \qquad \rho^2 - \theta_x - \theta_y > 0. \tag{28}$$

Together these are equation 19, and the roots are equation 20.

To classify the two roots we compute the Hessian of f at a stationary point. Using the stationarity condition in the form $D_u/D = (1-\theta_x)/(1+u)$,

$$\frac{\partial^2 f}{\partial u^2} = \frac{1}{2}\Big[-\frac{1-\theta_x}{(1+u)^2} + \frac{D_u^2}{D^2}\Big] = \frac{1}{2}\cdot\frac{1-\theta_x}{(1+u)^2}\big[(1-\theta_x) - 1\big] = -\frac{\theta_x(1-\theta_x)}{2(1+u)^2} < 0, \tag{29}$$

and similarly $\partial^2 f/\partial v^2 = -\theta_y(1-\theta_y)/[2(1+v)^2] < 0$. For the mixed derivative one finds, after expanding the products,

$$\frac{\partial^2 f}{\partial u\,\partial v} = \frac{D_u D_v - (1-\rho^2)D}{2D^2} = \frac{\rho^2}{2D^2}, \tag{30}$$

because $D_u D_v - (1-\rho^2)D = \rho^2$ identically. The diagonal entries are negative at both roots, so a root is a local maximum if the determinant of the Hessian is positive and a saddle if it is negative. Substituting $\theta_x = \rho^2 v/D$ and $\theta_y = \rho^2 u/D$ from equation 18, the determinant works out to

$$\det H = \frac{\rho^4}{4D^3(1+u)(1+v)}\Big[(1-\rho^2)\,uv - 1\Big]. \tag{31}$$

So the sign question reduces to whether $(1-\rho^2)uv$ is larger or smaller than 1. Now regard the right-hand side of the quadratic as a function of D along the line $u = \theta_y D/\rho^2$, $v = \theta_x D/\rho^2$:

$$G(D) := 1+u+v+(1-\rho^2)uv - D, \qquad G'(D) = \frac{u+v+2(1-\rho^2)uv}{D} - 1 = \frac{(1-\rho^2)uv - 1}{D}, \tag{32}$$

where the last step uses $D = 1+u+v+(1-\rho^2)uv$. G is a convex parabola with roots $D_- < D_+$, so $G'(D_+) > 0$ and $G'(D_-) < 0$. Hence $(1-\rho^2)uv > 1$ at D_+ (a local maximum) and < 1 at D_- (a saddle). At the spinodal the two roots merge and $(1-\rho^2)u^*v^* = 1$. Combined with $u/v = \theta_y/\theta_x$ this is equation 22.

(iii) The global maximum. Since $(S_X; S_Y)$ is bounded by $-\frac{1}{2}\log(1-\rho^2)$ while the compression costs grow without bound, $f \to -\infty$ as $u \to \infty$ or $v \to \infty$, so the maximum over $u, v \geq 0$ is attained. On the boundary faces $u = 0$ or $v = 0$ the shared term vanishes and $f \leq 0$, with equality only at the origin. In the interior the only critical points are the saddle D_- and the local maximum D_+. The global maximum is therefore either the origin, with value 0, or D_+, with value $f(u_+, v_+)$. Rearranging $f(u_+, v_+)$ using $\log[(1+u)(1+v)/D] - \theta_x\log(1+u) - \theta_y\log(1+v) = (1-\theta_x)\log(1+u) + (1-\theta_y)\log(1+v) - \log D$ gives equation 21.