Generalized Information Bottleneck for Gaussian Variables

Anonymous authors
Paper under double-blind review

Abstract

The information bottleneck (IB) method offers an attractive framework for understanding representation learning, however its applications are often limited by its computational intractability. Analytical characterization of the IB method is not only of practical interest, but it can also lead to new insights into learning phenomena. Here we consider a generalized IB problem, in which the mutual information in the original IB method is replaced by correlation measures based on Rényi and Jeffreys divergences. We derive an exact analytical IB solution for the case of Gaussian correlated variables. Our analysis reveals a series of structural transitions, similar to those previously observed in the original IB case. We find further that although solving the original, Rényi and Jeffreys IB problems yields different representations in general, the structural transitions occur at the same critical tradeoff parameters, and the Rényi and Jeffreys IB solutions perform well under the original IB objective. Our results suggest that formulating the IB method with alternative correlation measures could offer a strategy for obtaining an approximate solution to the original IB problem.

1 Information Bottleneck

Effective representation of data is key to generalizable learning. Characterizing what makes such representation good and how it emerges is crucial to understanding the success of modern machine learning. The information bottleneck (IB) method—an information-theoretic formulation for representation learning (Tishby et al., 1999)—has proved a particularly useful conceptual framework for this question, and has led to a deeper understanding of representation learning in both supervised and self-supervised learning (Achille & Soatto, 2018; Achille & Soatto, 2018; Tian et al., 2020; Zbontar et al., 2021). Investigating this abstraction of representation learning has the potential to yield new insights that are applicable to learning problems.

Quantifying the goodness of a representation requires the knowledge of what is to be learned from data. Information bottleneck theory exploits the fact that, in many settings, we can define relevant information through an additional variable; for example, it could be the label of each image in a classification task. This notion of relevance allows for a precise definition of optimality—an IB optimal representation T is maximally predictive of the relevance variable Y while minimizing the number of bits extracted from the data X. The IB method formulates this principle as an optimization problem (Tishby et al., 1999),

$$\min_{Q_{T|X}} \; I(T;X) - \beta I(T;Y). \tag{1}$$

Here the optimization is over the encoders $Q_{T|X}$ which provide a (stochastic) mapping from X to T. Maximizing the mutual information $I(T;Y)$ [second term in Eq (1)] encourages a representation T to encode more relevant information while minimizing $I(T;X)$ [first term in Eq (1)] discourages it from encoding irrelevant bits. The parameter $\beta > 0$ controls the fundamental tradeoff between the two information terms.

The IB method has proved successful in a number of applications, including neural coding (Palmer et al., 2015; Wang et al., 2021), statistical physics (Still et al., 2012; Gordon et al., 2021; Kline & Palmer, 2022), clustering (Strouse & Schwab, 2019), deep learning (Alemi et al., 2017; Achille & Soatto, 2018; Achille & Soatto, 2018), reinforcement learning (Goyal et al., 2019) and learning theory (Bialek et al., 2001; Shamir et al., 2010; Bialek et al., 2020; Ngampruetikorn & Schwab, 2022). However the nonlinear nature of the IB

problem makes it computationally costly. Although scalable learning methods based on the IB principle are possible thanks to variational bounds of mutual information (Alemi et al., 2017; Chalk et al., 2016; Poole et al., 2019), the choice of such bounds, as well as specific details on their implementations, can introduce strong inductive bias that competes with the original objective (Tschannen et al., 2020).

While large-scale applications of the IB method in its exact form are generally intractable, special cases exist. For example, the limit of low information—i.e., when both terms in Eq (1) are small—can be described by a perturbation theory, which provides a recipe for identifying a representation that yields maximum relevant information per extracted bit (Wu et al., 2019; Ngampruetikorn & Schwab, 2021). But perhaps the most important special case is when the source X and the target Y are Gaussian correlated random variables. In this case, an exact *analytical* solution exists (Chechik et al., 2005).

Although originally formulated with Shannon mutual information, the fundamental tradeoff in the IB method applies more generally: the IB optimization, Eq (1), remains well-defined when the information terms are replaced by appropriate mutual dependence measures. In this work, we consider generalized IB problems based on two important correlation measures. The first is a parametric generalization of Shannon information, based on Rényi divergence (Rényi, 1961). Rényi-based generalizations of mutual information and entropy are central in quantifying quantum entanglement (Horodecki et al., 2009; Eisert et al., 2010) and have proved a powerful tool in Monte-Carlo simulations (Hastings et al., 2010; Singh et al., 2011; Herdman et al., 2017) as well as in experiments (Islam et al., 2015; Bergschneider et al., 2019; Brydges et al., 2019). The second mutual dependence measure we consider is based on Jeffreys divergence (Jeffreys, 1946). The resulting Jeffreys information is (up to a constant prefactor) equal to the generalization gap of a broad family of learning algorithms, known as Gibbs algorithms (Aminian et al., 2021).

We derive an analytical IB solution for the case in which X and Y are Gaussian correlated, generalizing the result of Chechik et al. (2005) to a class of information-theoretic mutual dependence measures which includes Shannon information as a limiting case. We show that, for both Rényi and Jeffreys cases, an optimal encoder can be constructed from the eigenmodes of the normalized regression matrix $\Sigma_{X|Y}\Sigma_X^{-1}$. Our solution reveals a series of phase transitions, similar to those observed in the Gaussian IB method (Chechik et al., 2005). In both Rényi and Jeffreys cases, we find that although the optimal encoders depend on information measures, the phase transitions occur at the critical tradeoff parameters $\beta_c^{(i)}$ that coincide with that of the Shannon case, independent of the order of Rényi information.

2 Divergence-based Correlation measure

When two random variables X and Y are uncorrelated, their joint distribution P_{XY} is equal to the product of their marginals P_X and P_Y. As a result, we can quantify the mutual dependence between X and Y by the difference between P_{XY} and $P_X \otimes P_Y$,

$$\Omega(X;Y) \equiv \mathcal{D}(P_{XY} \parallel P_X \otimes P_Y) \geq 0. \tag{2}$$

Here $\mathcal{D}(P \parallel Q)$ denotes a statistical divergence which, by definition, is nonnegative and vanishes if and only if $P = Q$. When defined with the Kullback–Leibler (KL) divergence, the above measure becomes Shannon information, $I(X;Y) = D_{\mathrm{KL}}(P_{XY} \parallel P_X \otimes P_Y)$.

3 Rényi q–information

We consider a correlation measure, based on Rényi divergence (Rényi, 1961). More precisely, we define *Rényi q-information* as

$$I_q(X;Y) \equiv \mathcal{R}_q(P_{XY} \parallel P_X \otimes P_Y), \tag{3}$$

where $\mathcal{R}_q$ denotes Rényi divergence of order q,

$$\mathcal{R}_q(P \parallel Q) = \frac{1}{q-1} \ln \int dQ \left(\frac{dP}{dQ}\right)^q \tag{4}$$

for $q \in (0,1) \cup (1,\infty)$. This definition extends to $q = 0, 1$ and ∞ via continuity in q. In particular, $\mathcal{R}_1(P\|Q) = D_{\mathrm{KL}}(P\|Q)$ (van Erven & Harremoës, 2014, Thm 5), and as a result $I_1(X;Y) = I(X;Y)$. Rényi divergences, and thus q-information, satisfy the data processing inequality since they have a strictly increasing relationship with an f-divergence [with $f(t) = (t^q - 1)/(q-1)$] which exhibits this property, see, e.g., Liese & Vajda (2006).

3.1 Gaussian variables

For Gaussian correlated variables X and Y, the q-information is given by (see Appendix B for derivation)

$$I_q(X;Y) = -\frac{1}{2\bar{q}} \ln \frac{|\Sigma_{X|Y}\Sigma_X^{-1}|^{\bar{q}}}{|I - \bar{q}^2(I - \Sigma_{X|Y}\Sigma_X^{-1})|} \quad \text{with} \quad \bar{q} = 1 - q, \tag{5}$$

where I denotes the identity matrix in compatible dimensions. We see that this information depends on the covariance matrices only through the normalized regression matrix $\Sigma_{X|Y}\Sigma_X^{-1}$. We note also that this information can diverge when $q > 2$ since the eigenvalues of $\Sigma_{X|Y}\Sigma_X^{-1}$ range from zero to one (Chechik et al., 2005, Lemma B.1). It is easy to verify that Shannon information corresponds to the limit $q \to 1$,

$$I(X;Y) = \lim_{q\to 1} I_q(X;Y) = -\frac{1}{2} \ln |\Sigma_{X|Y}\Sigma_X^{-1}|. \tag{6}$$

In addition, we note that for Gaussian variables $I_2(X;Y) = 2I(X;Y)$ and $I_q(X;Y)$ increases with q from zero at $q = 0$.

Note that alternative definitions of Rényi mutual information exist. In physics literature, a frequently used definition is $I_q(X;Y) = S_q(X) + S_q(Y) - S_q(X,Y)$ where $S_q(X) = (1-q)^{-1} \ln \int dx\, p_X(x)^q$ is Rényi (differential) entropy of order q. However, for Gaussian variables, this definition leads to Rényi information that is equal to Shannon information regardless of q; the resulting Rényi IB problem is therefore identical to the original IB problem.

4 Rényi Information Bottleneck for Gaussian variables

Replacing the mutual information in the original IB objective [Eq (1)] with q-information yields

$$\mathcal{L}_q[Q_{T|X}] = I_q(T;X) - \beta I_q(T;Y) \tag{7}$$

where X denotes the source data, Y the target variable and T the representation of X. The loss function varies with the encoder $Q_{T|X}$ which provides a stochastic mapping from X to T. In general, the q-information terms need not be of the same order but the data processing inequality $I_q(T;X) \geq I_{q'}(T;Y)$ is guaranteed only when $q = q'$.

We specialize to the case where X and Y are Gaussian correlated and consider a family of noisy linear encoders,

$$T = AX + \xi \quad \text{with} \quad \xi \sim N(0, \Sigma_\xi). \tag{8}$$

Since Rényi divergences are invariant under an invertible transformation of random variables [see Eq (4)], we can transform T such that Σ_ξ becomes the identity matrix without changing the information content. In the following analysis, we set $\Sigma_\xi = I$ without loss of generality. That is, the encoder becomes a Gaussian channel, parametrized only by the matrix A,

$$T \mid X \sim N(AX, I). \tag{9}$$

To compute the information in Eq (7), we first marginalize out X from the above equation, yielding

$$T \sim N(A\mu_X, I + A\Sigma_X A^\mathsf{T}) \tag{10}$$

$$T \mid Y \sim N(A\mu_{X|Y}, I + A\Sigma_{X|Y} A^\mathsf{T}), \tag{11}$$

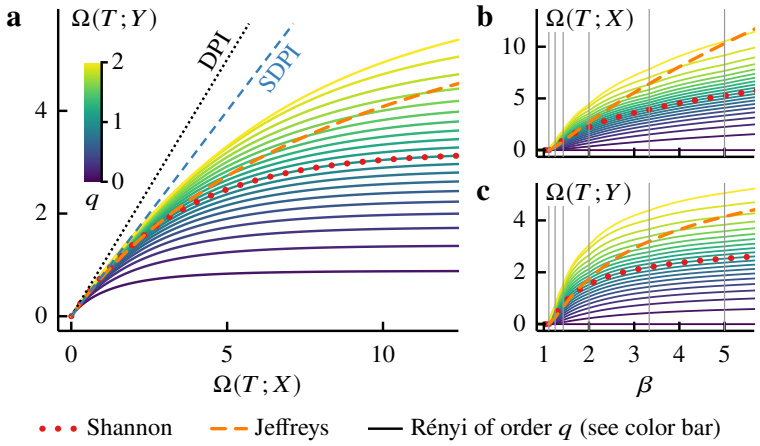

Figure 1: **a** The optimal frontiers of the generalized IB methods based on generalized correlation measures $\Omega(A;B)$, including Shannon ($\Omega = I$), Jeffreys ($\Omega = J$) and Rényi ($\Omega = I_q$) informations (see legend). For the Rényi case, we depict the results for a range of Rényi orders q (see color bar). We emphasize that while Shannon information is equivalent to Rényi information of order one ($q = 1$), Jeffreys information is not a special case of Rényi information. The relevant information $\Omega(T;Y)$ is bounded by the data processing inequality (DPI, black dotted line), $\Omega(T;Y) \leq \Omega(T;X)$. We also depict the tight, data-dependent version of DPI—the strong data processing inequality (SDPI, blue dahsed line), $\Omega(T;Y) \leq (1-\lambda_{\min})\Omega(T;X)$, where $\lambda_{\min}$ is the smallest eigenvalue of the normalized regression matrix $\Sigma_{X|Y}\Sigma_X^{-1}$. Note that the DPI and SDPI are the same for all information measures shown. **b-c** The extracted and relevant bits, $\Omega(T;X)$ and $\Omega(T;Y)$ respectively, increase with the tradeoff parameter β and vanish below the critical value $\beta_c = 1/(1-\lambda_{\min})$. The vertical lines mark the location of the critical tradeoff parameters (Eqs (22) & (35)). Here the eigenvalues of $\Sigma_{X|Y}\Sigma_X^{-1}$ are $\lambda_i = 0.1, 0.2, 0.3, 0.5, 0.7, 0.8$.

where we use $X \sim N(\mu_X, \Sigma_X)$ and $X\,|\,Y \sim N(\mu_{X|Y}, \Sigma_{X|Y})$. Substituting the covariance matrices in the above equations into Eq (5) results in

$$I_q(T;X) = -\frac{1}{2\bar{q}} \ln \frac{|I + A\Sigma_X A^\mathsf{T}|^q}{|I + (1-\bar{q}^2)A\Sigma_X A^\mathsf{T}|} \tag{12}$$

$$I_q(T;Y) = -\frac{1}{2\bar{q}} \ln \frac{|I + A\Sigma_{X|Y} A^\mathsf{T}|^{\bar{q}} |I + A\Sigma_X A^\mathsf{T}|^q}{|I + A[I - \bar{q}^2(I - \Sigma_{X|Y}\Sigma_X^{-1})]\Sigma_X A^\mathsf{T}|}. \tag{13}$$

Following the analysis of Chechik et al. (2005), we define the *mixing matrix* W such that

$$A = WV, \tag{14}$$

where V is a matrix of left (row) eigenvectors of the normalized regression matrix $\Sigma_{X|Y}\Sigma_X^{-1}$, i.e.,

$$V\Sigma_{X|Y}\Sigma_X^{-1} = \Lambda V \quad \text{with} \quad \Lambda = \text{diag}(\lambda_1, \lambda_2, \cdots). \tag{15}$$

We note that $V\Sigma_X^{1/2}$ is orthogonal and thus $V\Sigma_X V^\mathsf{T}$ is a diagonal matrix (Chechik et al., 2005, Lemma B.1), i.e.,

$$V\Sigma_X V^\mathsf{T} = R \quad \text{with} \quad R = \text{diag}(r_1, r_2, \cdots). \tag{16}$$

Writing Eqs (12-13) in terms of W, Λ and R leads to

$$I_q(T;X) = -\frac{1}{2\bar{q}} \ln \frac{|I + WRW^\mathsf{T}|^q}{|I + (1-\bar{q}^2)WRW^\mathsf{T}|} \tag{17}$$

$$I_q(T;Y) = -\frac{1}{2\bar{q}} \ln \frac{|I + W\Lambda RW^\mathsf{T}|^{\bar{q}} |I + WRW^\mathsf{T}|^q}{|I + W[I - \bar{q}^2(I-\Lambda)]RW^\mathsf{T}|}. \tag{18}$$

Substituting Eqs (17-18) into Eq (7) and setting its first order derivative with respect to the mixing matrix W to zero yields the first order condition

$$\frac{q}{I+RW^\mathsf{T}W}-\frac{1-\bar{q}^2}{I+(1-\bar{q}^2)RW^\mathsf{T}W}=\beta\bigg(\frac{q}{I+RW^\mathsf{T}W}+\frac{\bar{q}}{I+\Lambda RW^\mathsf{T}W}\Lambda \\ -\frac{1}{I+[I-\bar{q}^2(I-\Lambda)]RW^\mathsf{T}W}[I-\bar{q}^2(I-\Lambda)]\bigg). \tag{19}$$

In deriving the above, we use the identity

$$\frac{d}{dW}\ln|I+WCW^\mathsf{T}|=2W(I+CW^\mathsf{T}W)^{-1}C$$

which holds for any compatible square matrix C. We also assume that R and W are invertible.

We seek a solution of the form

$$RW^\mathsf{T}W=\mathrm{diag}(r_1w_1^2,r_2w_2^2,\cdots)\equiv\mathrm{diag}(u_1,u_2,\cdots). \tag{20}$$

Substituting this ansatz into Eq (19) results in

$$\frac{1}{\beta}=g_q(u_i,\lambda_i)=\frac{1-\lambda_i}{1+u_i\lambda_i}\frac{1+\frac{\bar{q}(1+\bar{q})(1-\lambda_i)u_i}{1+(1-\bar{q}^2(1-\lambda_i))u_i}}{1+\frac{\bar{q}(1+\bar{q})u_i}{1+(1-\bar{q}^2)u_i}}. \tag{21}$$

We see that the contributions from the eigenmodes of $\Sigma_{X|Y}\Sigma_X^{-1}$ decouple from one another and the reduced mixing weight, $u_i=r_iw_i^2$, for each mode depends only on the eigenvalue of that mode λ_i, the IB tradeoff parameter β and the order of Rényi information q. For $\lambda\in(0,1)$ and $q\in[0,2]$, the function $g_q(u,\lambda)$ is strictly decreasing in u for $u\geq 0$ and approaches zero as $u\to\infty$ (see Appendix A). As a result, Eq (21) has exactly one positive solution $u_i>0$ when $1/\beta<g_q(0,\lambda_i)$. That is, the eigenmode with eigenvalue λ_i contributes to the Rényi IB encoder only when β exceeds the critical value

$$\beta_c^{(i)}=\frac{1}{g_q(0,\lambda_i)}=\frac{1}{1-\lambda_i}. \tag{22}$$

Note that $\beta_c^{(i)}$ does not depend on q. To obtain u_i, we can either directly solve Eq (21) or use the analytical formula for the roots of the equivalent cubic equation (omitting the eigenmode indices)

$$0=au^3+bu^2+cu+d, \tag{23}$$

where the coefficients are given by

$$\begin{aligned}
a&=\lambda(1+\bar{q})(1-(1-\lambda)\bar{q}^2)\\
b&=\lambda(2+2\bar{q}+\lambda\bar{q}^2)+d(1+\bar{q})(1-\lambda\bar{q}-(1-\lambda)\bar{q}^2)\\
c&=\lambda(1+\bar{q}+\bar{q}^2)+d(2+(1-\lambda)\bar{q}-\bar{q}^2)\\
d&=1-\beta(1-\lambda).
\end{aligned}$$

Although the above calculation does not uniquely determine the mixing matrix W, we can obtain a valid IB encoder by taking $W=\mathrm{diag}(w_1,w_2,\cdots)$ where $w_i=\sqrt{u_i/r_i}$ since the Rényi-IB loss depends on W only through the diagonal entries of $W^\mathsf{T}W$. To see this, we substitute Eq (20) into Eqs (17-18) and write down

$$I_q(T;X)=-\frac{1}{2\bar{q}}\sum_i^{\beta>\beta_c^{(i)}}\ln\frac{(1+u_i)^q}{1+(1-\bar{q}^2)u_i} \tag{24}$$

$$I_q(T;Y)=-\frac{1}{2\bar{q}}\sum_i^{\beta>\beta_c^{(i)}}\ln\frac{(1+\lambda_iu_i)^{\bar{q}}(1+u_i)^q}{1+[1-\bar{q}^2(1-\lambda_i)]u_i}, \tag{25}$$

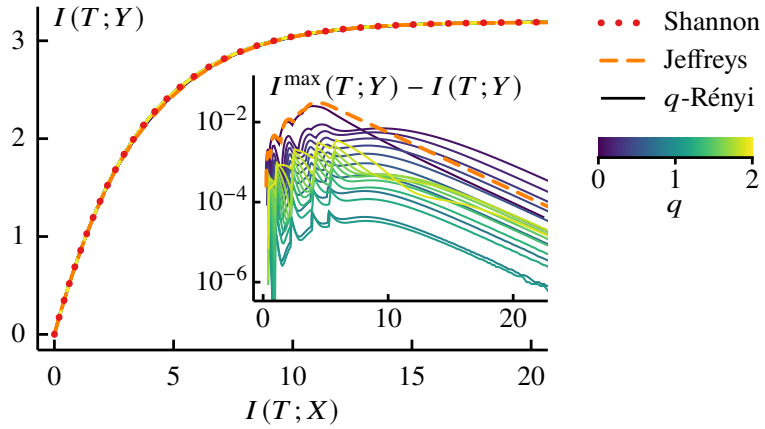

Figure 2: Solving Rényi and Jeffreys IB problems yields representations that are close to Shannon IB optimal. Plotted on the Shannon information plane, the solutions to Shannon (dotted), Jeffreys (dashed) and Rényi (solid) IB problems are nearly indistinguishable. For the Rényi case, we depict the results for a range of Rényi orders q (see color bar). *Inset:* We depict the gap between the maximum achievable and encoded relevant Shannon informations, $I^{\max}(T;Y)$ and $I(T;Y)$ respectively, as a function of the extracted Shannon information $I(T;X)$. This gap vanishes in the low and high-information limits, $I(T;X) \to 0$ and $I(T;X) \to \infty$. The satellite peaks result from the fact that the solutions to Shannon, Jeffreys and Rényi IB problems go through structural transitions at different values of $I(T;X)$ even though these transitions occur at the same set of critical tradeoff parameters. Here the eigenvalues of $\Sigma_{X|Y}\Sigma_X^{-1}$ are $\lambda_i = 0.1, 0.2, 0.3, 0.5, 0.7, 0.8$.

where the summations are restricted to the eigenmodes that contribute the IB encoder, i.e., those with $u_i > 0$. We depict the optimal frontiers of Rényi IB in Fig 1.

To complete our analysis of Rényi IB, we note that the analytical solution of Chechik et al. (2005) is a limiting case of our results. In the limit $q \to 1$, Eq (21) reads

$$\frac{1}{\beta} = g_{q=1}(u_i, \lambda_i) = \frac{1-\lambda_i}{1+u_i\lambda_i} \implies u_i^{(q=1)} = \frac{\beta(1-\lambda_i)-1}{\lambda_i}. \tag{26}$$

Recalling that $u_i = r_i w_i^2$, we see immediately that this solution is identical to that in Chechik et al. (2005, Lemma 4.1).

5 Jeffreys Information Bottleneck for Gaussian variables

The technique in the previous section applies also to the IB problems, based on other statistical divergences. In this section, we consider Jeffreys IB, defined by the loss function

$$\mathcal{L}_{\mathrm{J}}[Q_{T|X}] = J(T;X) - \beta J(T;Y). \tag{27}$$

Here $J(X;Y)$ denotes Jeffreys information which is a mutual dependence measure, defined by

$$J(X;Y) \equiv D_{\mathrm{J}}(P_{XY} \parallel P_X \otimes P_Y), \tag{28}$$

where D_{J} is Jeffreys divergence (Jeffreys, 1946),

$$D_{\mathrm{J}}(P \parallel Q) = \frac{1}{2}[D_{\mathrm{KL}}(P \parallel Q) + D_{\mathrm{KL}}(Q \parallel P)]. \tag{29}$$

For Gaussian correlated random variables, Jeffreys information takes a simple form (see Appendix C for derivation)

$$J(X;Y) = \frac{1}{2}\,\mathrm{tr}\left(\Sigma_X\Sigma_{X|Y}^{-1} - I\right). \tag{30}$$

Using the linear encoder from Eq (9), the information terms in Eq (27) read

$$J(T;X) = \frac{1}{2}\,\mathrm{tr}\left(WRW^{\mathsf{T}}\right) \tag{31}$$

$$J(T;Y) = \frac{1}{2}\,\mathrm{tr}\left((I + WRW^{\mathsf{T}})\frac{1}{I + W\Lambda RW^{\mathsf{T}}} - I\right). \tag{32}$$

where W, Λ and R are defined in Eqs (14-16).

To solve the IB optimization, we differentiate of the loss function with respect to the mixing matrix W and set the resulting derivative to zero, yielding

$$I = \beta\frac{1}{I + \Lambda RW^{\mathsf{T}}W}\left(I - (I + RW^{\mathsf{T}}W)\frac{1}{I + \Lambda RW^{\mathsf{T}}W}\Lambda\right), \tag{33}$$

where we use the identities

$$\frac{\partial A^{-1}}{\partial a} = -A^{-1}\frac{\partial A}{\partial a}A^{-1} \quad \text{and} \quad \frac{\partial}{\partial A}\,\mathrm{tr}(CABA^{\mathsf{T}}) = 2CAB$$

which hold for symmetric matrices B and C. We see that this equation is solvable by taking $W^{\mathsf{T}}W$ to be diagonal. Substituting Eq (20) into the above equation and solving for u_i gives

$$u_i = \frac{\sqrt{\beta(1-\lambda_i)} - 1}{\lambda_i}. \tag{34}$$

Since $u_i = r_i w_i^2 \geq 0$, we see that this solution is valid only when the tradeoff parameter β in greater than the critical value

$$\beta_c^{(i)} = \frac{1}{1-\lambda_i}. \tag{35}$$

We note that this critical value is identical to that of Rényi IB [Eq (22)]. Finally substituting Eq (34) into Eqs (31-32) leads to

$$J(T;X) = \frac{1}{2}\sum_i^{\beta>\beta_c^{(i)}} \frac{\sqrt{\beta(1-\lambda_i)} - 1}{\lambda_i} \tag{36}$$

$$J(T;Y) = \frac{1}{2}\sum_i^{\beta>\beta_c^{(i)}} \frac{1-\lambda_i}{\lambda_i}\frac{\sqrt{\beta(1-\lambda_i)} - 1}{\sqrt{\beta(1-\lambda_i)}}. \tag{37}$$

where the summations are limited to the modes that contribute to the encoder, i.e., those with $\beta_c^{(i)} < \beta$. In Fig 1, we depict an example of the Jeffreys IB optimal frontier, computed from the above equations. We emphasize that while Shannon information is equivalent to Rényi information with $q = 1$, Jeffreys information is not a special case of Rényi information.

6 Discussion & Conclusion

In Fig 2, we depict the solutions to the original, Rényi and Jeffreys IB problems on the Shannon information plane. We see that these solutions are very close to the optimal frontier, characterized by the Shannon IB solutions. This result suggests that formulating and solving an IB problem, defined with alternative correlation measures other than Shannon information, could offer a strategy for obtaining an approximate solution to the original IB problem. To better illustrate the differences between the solutions to the original, Rényi and Jeffreys IB problems, the inset shows how much less relevant Shannon information the optimal representations of Rényi and Jeffreys IB encode, compared to the Shannon IB optimal representation. We see that the differences are maximum at intermediate information and vanish in the low and high-information limits. In addition, the Shannon information gaps exhibit satellite peaks, resulting from structural the

transition of the IB solutions. We note that although these transitions occur at the same critical tradeoff parameters [Eqs (22) & (35)], they generally correspond to different values of extracted Shannon information.

To sum up, we consider generalized IB problems in which the mutual information is replaced by mutual dependence measures, based on Rényi and Jeffreys divergences. We obtain exact analytical solutions for the case of Gaussian correlated random variables, generalizing the results of Chechik et al. (2005). We show that the fundamental IB tradeoff between relevance and compression holds also for correlation measures other than Shannon information. Our analyses reveal structural transitions in the optimal representations, similar to that in the original IB method (Chechik et al., 2005). Interestingly the critical tradeoff parameters are the same for original, Rényi and Jeffreys IB problems, even though the solutions are distinct.

We anticipate that our work will find application in physics of correlated components which relies on Rényi-generalization of entropy and information to quantify entanglement. In addition, our characterization of Jeffreys IB could have implications for understanding the generalization properties of Gibbs learning algorithms of which the generalization gap is proportional to Jeffreys information between fitted models and training data. Finally we note that the conditional IB problem, in which the compression term $I(T;X)$ is replaced by $I(T;X\,|\,Y)$, becomes non-trivial for generalized information measures since the chain rule does not hold for Rényi and Jeffreys information—that is, given the Markov constraint T–X–Y, we have $I(T;X\,|\,Y) = I(T;X) - I(T;Y)$ for Shannon information, but in general, $\Omega(T;X\,|\,Y) \neq \Omega(T;X) - \Omega(T;Y)$. The logical steps in our analyses are readily generalizable to conditional IB problems.

References

A. Achille and S. Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2897–2905, 2018. doi:10.1109/TPAMI.2017.2784440.

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018. URL http://jmlr.org/papers/v19/17-646.html.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HyxQzBceg.

Gholamali Aminian, Yuheng Bu, Laura Toni, Miguel Rodrigues, and Gregory Wornell. An exact characterization of the generalization error for the gibbs algorithm. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 8106–8118. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/445e24b5f22cacb9d51a837c10e91a3f-Abstract.html.

Andrea Bergschneider, Vincent M. Klinkhamer, Jan Hendrik Becher, Ralf Klemt, Lukas Palm, Gerhard Zürn, Selim Jochim, and Philipp M. Preiss. Experimental characterization of two-particle entanglement through position and momentum correlations. *Nature Physics*, 15(7):640–644, 2019. doi:10.1038/s41567-019-0508-6.

William Bialek, Ilya Nemenman, and Naftali Tishby. Predictability, Complexity, and Learning. *Neural Computation*, 13(11):2409–2463, 11 2001. doi:10.1162/089976601753195969.

William Bialek, Stephanie E. Palmer, and David J. Schwab. What makes it possible to learn probability distributions in the natural world?, 2020.

Tiff Brydges, Andreas Elben, Petar Jurcevic, Benoît Vermersch, Christine Maier, Ben P. Lanyon, Peter Zoller, Rainer Blatt, and Christian F. Roos. Probing Rényi entanglement entropy via randomized measurements. *Science*, 364(6437):260–263, 2019. doi:10.1126/science.aau4963.

Matthew Chalk, Olivier Marre, and Gasper Tkacik. Relevant sparse codes with variational information bottleneck. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural*

Information Processing Systems, volume 29, pp. 1957–1965. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/hash/a89cf525e1d9f04d16ce31165e139a4b-Abstract.html.

Gal Chechik, Amir Globerson, Naftali Tishby, and Yair Weiss. Information bottleneck for Gaussian variables. *Journal of Machine Learning Research*, 6:165–188, 2005. URL https://www.jmlr.org/papers/v6/chechik05a.html.

J. Eisert, M. Cramer, and M. B. Plenio. Colloquium: Area laws for the entanglement entropy. *Reviews of Modern Physics*, 82:277–306, Feb 2010. doi:10.1103/RevModPhys.82.277.

Amit Gordon, Aditya Banerjee, Maciej Koch-Janusz, and Zohar Ringel. Relevance in the renormalization group and in information theory. *Physical Review Letters*, 126:240601, Jun 2021. doi:10.1103/PhysRevLett.126.240601.

Anirudh Goyal, Riashat Islam, DJ Strouse, Zafarali Ahmed, Hugo Larochelle, Matthew Botvinick, Sergey Levine, and Yoshua Bengio. Transfer and exploration via the information bottleneck. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJg8yhAqKm.

Matthew B. Hastings, Iván González, Ann B. Kallin, and Roger G. Melko. Measuring renyi entanglement entropy in quantum monte carlo simulations. *Physical Review Letters*, 104:157201, Apr 2010. doi:10.1103/PhysRevLett.104.157201.

C. M. Herdman, P. N. Roy, R. G. Melko, and A. Del Maestro. Entanglement area law in superfluid ^{4}He. *Nature Physics*, 13(6):556–558, 2017. doi:10.1038/nphys4075.

Ryszard Horodecki, Paweł Horodecki, Michał Horodecki, and Karol Horodecki. Quantum entanglement. *Reviews of Modern Physics*, 81:865–942, Jun 2009. doi:10.1103/RevModPhys.81.865.

Rajibul Islam, Ruichao Ma, Philipp M. Preiss, M. Eric Tai, Alexander Lukin, Matthew Rispoli, and Markus Greiner. Measuring entanglement entropy in a quantum many-body system. *Nature*, 528(7580):77–83, 2015. doi:10.1038/nature15750.

Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946. doi:10.1098/rspa.1946.0056.

Adam G Kline and Stephanie E Palmer. Gaussian information bottleneck and the non-perturbative renormalization group. *New Journal of Physics*, 24(3):033007, mar 2022. doi:10.1088/1367-2630/ac395d.

F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006. doi:10.1109/TIT.2006.881731.

Vudtiwat Ngampruetikorn and David J Schwab. Perturbation theory for the information bottleneck. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 21008–21018. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/af8d9c4e238c63fb074b44eb6aed80ae-Abstract.html.

Vudtiwat Ngampruetikorn and David J Schwab. Information bottleneck theory of high-dimensional regression: relevancy, efficiency and optimality. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 9784–9796. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/3fbcfbc2b4009ae8dfa17a562532d123-Abstract-Conference.html.

Stephanie E. Palmer, Olivier Marre, Michael J. Berry, and William Bialek. Predictive information in a sensory population. *Proceedings of the National Academy of Sciences*, 112(22):6908–6913, 2015. doi:10.1073/pnas.1506855112.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5171–5180. PMLR, 2019. URL https://proceedings.mlr.press/v97/poole19a.html.

A Rényi. On measures of entropy and information. In Jerzy Neyman (ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 547–561, 1961.

Ohad Shamir, Sivan Sabato, and Naftali Tishby. Learning and generalization with the information bottleneck. *Theoretical Computer Science*, 411(29):2696–2711, 2010. doi:10.1016/j.tcs.2010.04.006. Algorithmic Learning Theory (ALT 2008).

Rajiv R. P. Singh, Matthew B. Hastings, Ann B. Kallin, and Roger G. Melko. Finite-temperature critical behavior of mutual information. *Physical Review Letters*, 106:135701, Mar 2011. doi:10.1103/PhysRevLett.106.135701.

Susanne Still, David A. Sivak, Anthony J. Bell, and Gavin E. Crooks. Thermodynamics of prediction. *Physical Review Letters*, 109:120604, 2012. doi:10.1103/PhysRevLett.109.120604.

DJ Strouse and David J. Schwab. The information bottleneck and geometric clustering. *Neural Computation*, 31(3):596–612, 2019. doi:10.1162/neco_a_01136.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6827–6839. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c2e5eaae9152079b9e95845750bb9ab-Abstract.html.

Naftali Tishby, Fernando C. N. Pereira, and William Bialek. The information bottleneck method. In B. Hajek and R. S. Sreenivas (eds.), *37th Allerton Conference on Communication, Control and Computing*, pp. 368–377. University of Illinois, 1999. URL http://arxiv.org/abs/physics/0004057.

Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkxoh24FPH.

Tim van Erven and Peter Harremoës. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014. doi:10.1109/TIT.2014.2320500.

Siwei Wang, Idan Segev, Alexander Borst, and Stephanie Palmer. Maximally efficient prediction in the early fly visual system may support evasive flight maneuvers. *PLOS Computational Biology*, 17(5):e1008965, 05 2021. doi:10.1371/journal.pcbi.1008965.

Tailin Wu, Ian Fischer, Isaac L. Chuang, and Max Tegmark. Learnability for the information bottleneck. *Entropy*, 21(10):924, 2019. doi:10.3390/e21100924.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 2021. URL https://proceedings.mlr.press/v139/zbontar21a.html.

A Supplementary figure

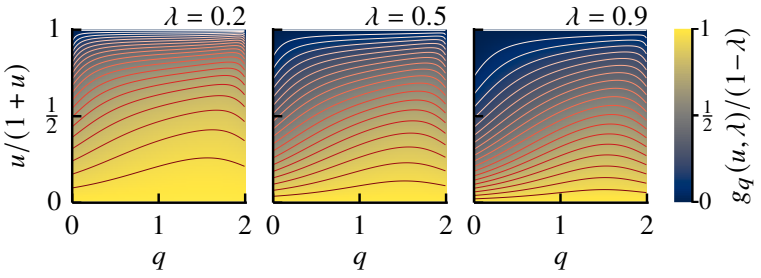

Figure A1: The function $g_q(u,\lambda)$ [Eq (21)] decreases with u from $1-\lambda$ at $u=0$ and approaches zero as $u\to\infty$. As a result, the equation $\beta^{-1}=g_q(u,\lambda)$ always has a unique positive solution when $\beta > 1/(1-\lambda)$. We consider only $0\le q\le 2$ since Rényi information for Gaussian variables can diverge for $q>2$ [see Eq (5)].

B Rényi information for Gaussian variables

In this appendix, we derive Rényi mutual information for Gaussian correlated variables. Using the definition from Eqs (3-4), we write down Rényi mutual information for continuous random variables,

$$I_q(X;Y)=\frac{1}{q-1}\ln\int dxdy\, p_X(x)p_Y(y)\left(\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}\right)^q. \tag{38}$$

where p_X, p_Y and p_{XY} denote the probability density functions of X, Y and (X,Y), respectively. We consider Gaussian correlated random variables

$$\begin{bmatrix}X\\Y\end{bmatrix}\sim N(\mu,\Sigma)\ \text{ with }\ \mu=\begin{bmatrix}\mu_X\\\mu_Y\end{bmatrix}\ \text{ and }\ \Sigma=\begin{bmatrix}\Sigma_X & \Sigma_{XY}\\ \Sigma_{YX} & \Sigma_Y\end{bmatrix}. \tag{39}$$

In this case, the joint probability density is given by

$$p_{XY}(x,y)=\frac{\exp\left\{-([\begin{smallmatrix}x\\y\end{smallmatrix}]-\mu)^\intercal\Sigma^{-1}([\begin{smallmatrix}x\\y\end{smallmatrix}]-\mu)\right\}}{|2\pi\Sigma|^{1/2}} \tag{40}$$

The product of the marginal distributions is equal to a joint distribution but with Σ_{XY} and Σ_{YX} set to zero, i.e.,

$$p_X(x)p_Y(y)=\frac{\exp\left\{-([\begin{smallmatrix}x\\y\end{smallmatrix}]-\mu)^\intercal\bar{\Sigma}^{-1}([\begin{smallmatrix}x\\y\end{smallmatrix}]-\mu)\right\}}{|2\pi\bar{\Sigma}|^{1/2}} \tag{41}$$

where $\bar{\Sigma}=\left[\begin{smallmatrix}\Sigma_X & \cdot\\ \cdot & \Sigma_Y\end{smallmatrix}\right]$. Substituting the above densities into Eq (38) and performing the resulting Gaussian integration over x and y gives

$$I_q(X;Y)=\frac{1}{q-1}\ln\frac{|q\Sigma^{-1}+(1-q)\bar{\Sigma}^{-1}|^{-1/2}}{|\Sigma|^{q/2}|\bar{\Sigma}|^{(1-q)/2}}. \tag{42}$$

The determinants of the covariance matrices are given by

$$|\Sigma|=|\Sigma_Y|\times|\Sigma_{X|Y}|\quad\text{and}\quad|\bar{\Sigma}|=|\Sigma_Y|\times|\Sigma_X|, \tag{43}$$

where $\Sigma_{X|Y}=\Sigma_X-\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX}$ and we use the identity

$$\begin{vmatrix}A & B\\ C & D\end{vmatrix}=|D|\times|A-BD^{-1}C|. \tag{44}$$

We now consider the numerator in Eq (42),

$$\begin{aligned}\left|q\Sigma^{-1}+(1-q)\bar{\Sigma}^{-1}\right| &= \left|\Sigma^{-1}\left(q\bar{\Sigma}+(1-q)\Sigma\right)\bar{\Sigma}^{-1}\right| \\ &= \frac{1}{|\Sigma|\times|\bar{\Sigma}|}\begin{vmatrix}\Sigma_X & (1-q)\Sigma_{XY}\\ (1-q)\Sigma_{YX} & \Sigma_Y\end{vmatrix} \\ &= \frac{\left|I-(1-q)^2(I-\Sigma_{X|Y}\Sigma_X^{-1})\right|}{|\Sigma_Y|\times|\Sigma_{X|Y}|},\end{aligned} \tag{45}$$

where the last equality follows from Eqs (43-44). Finally we write down the Rényi information for Gaussian variables

$$I_q(X;Y) = \frac{1/2}{q-1}\ln\frac{|\Sigma_{X|Y}\Sigma_X^{-1}|^{1-q}}{\left|I-(1-q)^2(I-\Sigma_{X|Y}\Sigma_X^{-1})\right|}. \tag{46}$$

This expression is identical to Eq (5) (with $\bar{q}=1-q$).

C Jeffreys information for Gaussian variables

The Jeffreys information is defined via

$$J(X;Y) \equiv D_{\mathrm{J}}(P_{XY} \parallel P_X \otimes P_Y), \tag{47}$$

where D_{J} is Jeffreys divergence (Jeffreys, 1946),

$$D_{\mathrm{J}}(P \parallel Q) = \frac{1}{2}[D_{\mathrm{KL}}(P \parallel Q) + D_{\mathrm{KL}}(Q \parallel P)]. \tag{48}$$

For Gaussian correlated X and Y, the Jeffreys information follows immediately from the KL divergence between two multivariate Gaussian distributions

$$\begin{aligned}D_{\mathrm{KL}}(N(\mu_0,\Sigma_0) \parallel N(\mu_1,\Sigma_1)) = \frac{1}{2}\Bigg(&\operatorname{tr}(\Sigma_1^{-1}\Sigma_0 - I) \\ &+ (\mu_1-\mu_0)^{\mathsf{T}}\Sigma_1^{-1}(\mu_1-\mu_0) + \ln\frac{|\Sigma_1|}{|\Sigma_0|}\Bigg).\end{aligned} \tag{49}$$

For X and Y described by Eq (39), we have $P_{XY}=N(\mu,\Sigma)$ and $P_X\otimes P_Y = N(\mu,\bar{\Sigma})$, where $\Sigma = \left[\begin{smallmatrix}\Sigma_X & \Sigma_{XY}\\ \Sigma_{YX} & \Sigma_Y\end{smallmatrix}\right]$ and $\bar{\Sigma} = \left[\begin{smallmatrix}\Sigma_X & \cdot\\ \cdot & \Sigma_Y\end{smallmatrix}\right]$. As a result, we have

$$D_{\mathrm{KL}}(P_{XY} \parallel P_X \otimes P_Y) = \frac{1}{2}\left(\operatorname{tr}(\bar{\Sigma}^{-1}\Sigma - I) + \ln\frac{|\bar{\Sigma}|}{|\Sigma|}\right) \tag{50}$$

$$D_{\mathrm{KL}}(P_X \otimes P_Y \parallel P_{XY}) = \frac{1}{2}\left(\operatorname{tr}(\Sigma^{-1}\bar{\Sigma} - I) + \ln\frac{|\Sigma|}{|\bar{\Sigma}|}\right). \tag{51}$$

We see that the logarithmic term drops out upon symmetrization [Eq (48)]. Substituting $\bar{\Sigma}^{-1} = \left[\begin{smallmatrix}\Sigma_X^{-1} & \cdot\\ \cdot & \Sigma_Y^{-1}\end{smallmatrix}\right]$ and the determinant formula in Eq (43) into Eq (50) gives

$$D_{\mathrm{KL}}(P_{XY} \parallel P_X \otimes P_Y) = \frac{1}{2}\ln\frac{|\bar{\Sigma}|}{|\Sigma|} = -\frac{1}{2}\ln|\Sigma_{X|Y}\Sigma_X^{-1}| \tag{52}$$

which is the usual mutual information, as expected. To compute trace in Eq (51), we write down the inverse of the covariance matrix,

$$\Sigma^{-1} = \begin{pmatrix}\Sigma_{X|Y}^{-1} & -\Sigma_{X|Y}^{-1}\Sigma_{XY}\Sigma_Y^{-1}\\ -\Sigma_Y^{-1}\Sigma_{YX}\Sigma_{X|Y}^{-1} & \Sigma_{Y|X}^{-1}\end{pmatrix}. \tag{53}$$

Therefore we have

$$\begin{aligned}
\operatorname{tr}(\Sigma^{-1}\bar{\Sigma} - I) &= \operatorname{tr}(\Sigma^{-1}(\bar{\Sigma} - \Sigma)) \\
&= \operatorname{tr}\left(\begin{bmatrix} \Sigma_{X|Y}^{-1} & -\Sigma_{X|Y}^{-1}\Sigma_{XY}\Sigma_{Y}^{-1} \\ -\Sigma_{Y}^{-1}\Sigma_{YX}\Sigma_{X|Y}^{-1} & \Sigma_{Y|X}^{-1} \end{bmatrix} \begin{bmatrix} \cdot & -\Sigma_{XY} \\ -\Sigma_{YX} & \cdot \end{bmatrix}\right) \\
&= \operatorname{tr}(\Sigma_{X|Y}^{-1}\Sigma_{XY}\Sigma_{Y}^{-1}\Sigma_{YX}) + \operatorname{tr}(\Sigma_{Y}^{-1}\Sigma_{YX}\Sigma_{X|Y}^{-1}\Sigma_{XY}) \\
&= 2\operatorname{tr}(\Sigma_{XY}\Sigma_{Y}^{-1}\Sigma_{YX}\Sigma_{X|Y}^{-1}) \\
&= 2\operatorname{tr}(\Sigma_{X}\Sigma_{X|Y}^{-1} - I),
\end{aligned} \tag{54}$$

where the last equality follows from the identity $\Sigma_{X|Y} = \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX}$. Substituting the above result into Eq (51) yields

$$D_{\mathrm{KL}}(P_X \otimes P_Y \parallel P_{XY}) = \operatorname{tr}(\Sigma_X\Sigma_{X|Y}^{-1} - I) + \frac{1}{2}\ln|\Sigma_{X|Y}\Sigma_X^{-1}|. \tag{55}$$

Finally eliminating the logarithmic term with Eq (52) leads to

$$\begin{aligned}
J(X;Y) &= \frac{1}{2}[D_{\mathrm{KL}}(P_X \otimes P_Y \parallel P_{XY}) + D_{\mathrm{KL}}(P_{XY} \parallel P_X \otimes P_Y)] \\
&= \frac{1}{2}\operatorname{tr}\left(\Sigma_X\Sigma_{X|Y}^{-1} - I\right).
\end{aligned} \tag{56}$$