On the Optimality of the Gaussian Information Bottleneck Curve

Amir Globerson and Naftali Tishby

School of Computer Science and Engineering and
The Interdisciplinary Center for Neural Computation
The Hebrew University of Jerusalem, 91904, Israel

February 20, 2004

1 Introduction

In a recent work [2], we address the following optimization problem, which we call the Gaussian Information Bottleneck: Consider two variables X, Y which are random vectors of dimensions n_x, n_y, and have a joint Gaussian distribution. We wish to find a variable T which is a compressed version of X, such that $I(T;X) = I_x$, and T is maximally informative about Y. This is cast as the following constrained optimization problem:

$$R(I_x) = \max_{p(t|x):T\to X\to Y, I(T;X)=I_x} I(T;Y) \tag{1}$$

The function $R(I_x)$ is called the information curve, and defines the maximum information one can keep about Y when forced to compress X by I_x.

In [2] we carry out this optimization under the limitation that $p(t|x)$ is Gaussian [1] . Denote by $G(T,X)$ the set of conditional Gaussian $p(t|x)$ distributions, then the Gaussian information curve is defined as

$$R^G(I_x) = \max_{p(t|x):T\to X\to Y, I(T;X)=I_x, p(t|x)\in G(T,X)} I(T;Y) \tag{2}$$

This function is calculated explicitly in [2] and will be described in what follows.

The main result of the current manuscript is that the optimization problems in Equations 1 and 2 are equivalent. This is stated below

Theorem 1 *$R(I_x) = R^G(I_x)$ and thus the maximum in Equation 1 can be achieved with a Gaussian $p(t|x)$*

[1] Which implies X, T are jointly Gaussian, since X itself is Gaussian

In [1] a similar problem is solved for one dimensional Gaussian variables, in the context of lossy coding with side information. Here we treat the multi-dimensional case, which has several interesting properties, and technical difficulties which are not present in the scalar case.

To prove the theorem, one needs to show that $R^G(I_x) \geq R(I_x)$ for all values of I_x. This is proven in the remainder of the text. The paper is organized as follows: Section 2 defines notations and characterizes certain covariance matrices. Section 3 describes the curve $R^G(I_x)$ which was calculated in [2]. Section 4 transforms the variables X, Y into a representation which is easier to work with. Finally, Section 5 gives the proof of Theorem 1.

2 Notations and matrix properties

In what follows we use various forms of covariance matrices, which we outline in this section.

Notation 1 *Denote by X_i the univariate Gaussian variable which is the i-th coordinate of the multivariate random variable X. We use Y_i with a similar notation.*

Notation 2 *Denote by $X^{(m,d)}$ the multivariate Gaussian of dimension d, $X^{(m,d)} \equiv (X_m, ..., X_{m+d-1})$. Also, denote $X^{(1,d)} \equiv X^{(d)}$.*

Notation 3 *Denote by Σ_{xy} the covariance matrix of X and Y. Thus the $(i,j)^{th}$ element of Σ_{xy} is $E(X_iY_j) - E(X_i)E(Y_j)$. The dimensions of Σ_{xy} are (n_x, n_y).*

Notation 4 *Denote by Σ_x the covariance matrix of X. Thus the $(i,j)^{th}$ element of Σ_x is $E(X_iX_j) - E(X_i)E(X_j)$. The dimensions of Σ_x are (n_x, n_x).*

Consider the variable $W = [X, Y]$ produced by concatenating X, Y. Its covariance matrix is thus given by

$$\Sigma_w = \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix}$$

The distribution $p(x|y)$ is known to be Gaussian as well. The covariance of this Gaussian does not depend on the value of y conditioned on. We denote the conditional covariance by $\Sigma_{x|y}$. Note that it is a square matrix of size n_x. The following identity expresses $\Sigma_{x|y}$ as a function of the unconditional covariances [3]

$$\Sigma_{x|y} = \Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx} \tag{3}$$

In what follows we make frequent use of the matrices $\Sigma_{x|y}\Sigma_x^{-1}$ and $\Sigma_{y|x}\Sigma_y^{-1}$. These are normalized conditional covariance matrices, which reflect by how much the knowledge of Y reduces the *spread* of X. For one dimensional variables the scalar $\Sigma_{x|y}\Sigma_x^{-1}$ will be zero if X, Y are deterministically related, and one if

X, Y are independent. In the multidimensional case, the values of interest are the eigenvalues of the matrices. As in the scalar case, eigenvalues close to zero reflect a deterministic relation between X and Y, and eigenvalues close to one reflect independence.

Notation 5 *Denote the eigenvalues of* $\Sigma_{x|y}\Sigma_x^{-1}$ *by* $\lambda_1, \ldots, \lambda_{n_x}$*, and assume they are sorted in an ascending order.*

We now prove some properties of these eigenvalues.

Lemma 1 *The eigenvalues of* $\Sigma_{x|y}\Sigma_x^{-1}$ *satisfy* $0 \le \lambda_i \le 1$ *for* $i = 1, \ldots, n_x$.

Proof:
The matrix $\Sigma_{x|y}\Sigma_x^{-1}$ can be written as

$$\Sigma_{x|y}\Sigma_x^{-1} = \left(\Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}\right)\Sigma_x^{-1} = I - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1}$$

The matrix $\Sigma_{x|y}\Sigma_x^{-1}$ has non-negative eigenvalues since its eigenvalues are identical to those of $\Sigma_x^{-\frac{1}{2}}\Sigma_{x|y}\Sigma_x^{-\frac{1}{2}}$ (see Section 6.3), which is Positive Semidefinite (PSD). Similarly, the matrix $\Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1}$ is PSD since its eigenvalues are identical to those of $\Sigma_x^{-\frac{1}{2}}\Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-\frac{1}{2}}$, which is also PSD. Thus $\Sigma_{x|y}\Sigma_x^{-1}$ has non-negative eigenvalues, but since it is equal to I minus a PSD matrix, its eigenvalues must also be smaller than one. □

The following Lemma states the eigenvalues of $\Sigma_{x|y}\Sigma_x^{-1}, \Sigma_{y|x}\Sigma_y^{-1}$ are essentially identical.

Lemma 2 *Assume* $n_x \ge n_y$. *Denote by* $\gamma_1, \ldots, \gamma_{n_y}$ *the sorted eigenvalues of* $\Sigma_{y|x}\Sigma_y^{-1}$. *Then* $\gamma_i = \lambda_i$ *for* $i = 1, \ldots, n_y$. *Furthermore* $\lambda_i = 1$ *for* $n_y < i \le n_x$

Proof:
Assume v is a left eigenvector of $\Sigma_{x|y}\Sigma_x^{-1}$ with eigenvalue λ, then

$$\begin{aligned}
v(I - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1}) &= \lambda v \\
v(\Sigma_{xy}\Sigma_y^{-1} - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy}\Sigma_y^{-1}) &= \lambda v\Sigma_{xy}\Sigma_y^{-1} \\
v\Sigma_{xy}\Sigma_y^{-1}(I - \Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy}\Sigma_y^{-1}) &= \lambda v\Sigma_{xy}\Sigma_y^{-1} \\
\left(v\Sigma_{xy}\Sigma_y^{-1}\right)\Sigma_{y|x}\Sigma_y^{-1} &= \lambda\left(v\Sigma_{xy}\Sigma_y^{-1}\right)
\end{aligned}$$

Thus $v\Sigma_{xy}\Sigma_y^{-1}$ is an eigenvector of $\Sigma_{y|x}\Sigma_y^{-1}$ with eigenvalue λ, and all the eigenvalues of $\Sigma_{x|y}\Sigma_x^{-1}$ are eigenvalues of $\Sigma_{y|x}\Sigma_y^{-1}$. which shows that there is a correspondence between the eigenvalue sets.

To see that $\lambda_i = 1$ for $n_y < i \le n_x$, note that $\Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1}$ has rank n_y at most, so it has an eigenvalue zero with multiplicity at least $n_x - n_y$. Thus the last $n_x - n_y$ eigenvalues of $\Sigma_{x|y}\Sigma_x^{-1}$ are equal to 1. □

3 The structure of $R^G(I_x)$

In [2] we derived the analytic form of $R^G(I_x)$. The matrix $\Sigma_{x|y}\Sigma_x^{-1}$ discussed in the previous section plays a central role in the derivation. We show in [2] that the curve $R^G(I_x)$ is made up of a set of segments, which form a continuous curve. In order to understand the functional form of the segments, we need the following definitions. Define $\hat{c}(k), \hat{n}(I)$ as the functions

$$
\begin{aligned}
\hat{c}(k) &= \begin{cases} 0 & k=1 \\ \sum_{i=1}^{k-1} \log \frac{\lambda_k}{\lambda_i} \frac{1-\lambda_i}{1-\lambda_k} & k>1 \end{cases} \\
\hat{n}(I) &= \arg \max_{1 \le k \le n_x} \hat{c}(k) \le I
\end{aligned}
$$

Note that $\hat{n}(I)$ partitions the values of I into n_x segments. In [2] we show that $R^G(I_x)$ has a slightly different form for each such segment, and is given by

$$
R^G(I_x) = I_x - \frac{\hat{n}(I_x)}{2} \log \left(\prod_{i=1}^{\hat{n}(I_x)} (1-\lambda_i)^{\frac{1}{\hat{n}(I_x)}} + e^{\frac{2I_x}{\hat{n}(I_x)}} \prod_{i=1}^{\hat{n}(I_x)} \lambda_i^{\frac{1}{\hat{n}(I_x)}} \right) \quad . \tag{4}
$$

The function $R^G(I_x)$ can be shown to be continuous and concave in I. An example of $R^G(I_x)$ is given in Figure 1, showing the points $I_x = \hat{c}(k)$ where $\hat{n}(I_x)$ switches values.

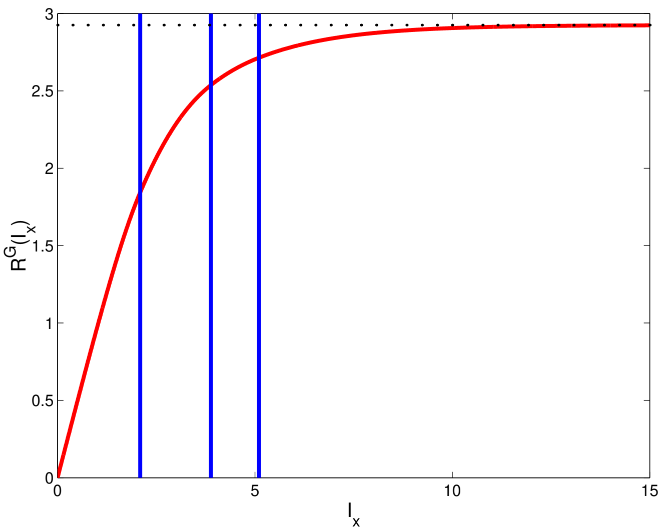

Figure 1: The curve $R^G(I_x)$ obtained with four eigenvalues λ_i=0.01,0.4,0.8,0.9. Values of I_x corresponding to $\hat{c}(2), \hat{c}(3), \hat{c}(4)$ (i.e where $\hat{n}(I_x)$ switches values) are marked by vertical lines. The asymptotic value of $I(X;Y)$ is marked by the horizontal line.

4 Canonical representation

Before proving the main theorem, we will transform the variables X and Y to a canonical representation, which is easier to manipulate. This is done via the eigenvectors of the matrices $\Sigma_{x|y}\Sigma_x^{-1}$ and $\Sigma_{y|x}\Sigma_y^{-1}$

Notation 6 *Denote A_x the eigenvector matrix of $\Sigma_{x|y}\Sigma_x^{-1}$ and B_y that of $\Sigma_{y|x}\Sigma_y^{-1}$.*

Define $\tilde{X} \equiv A_x X$ and $\tilde{Y} \equiv B_y Y$. The following Lemma gives some useful properties of the new variables (see proof in section 6.3)

Lemma 3 *The matrices A_x, B_y can be scaled so that the following properties hold*

1. *A_x, B_y are full rank*
2. *$\tilde{X}$ and $\tilde{Y}$ are spherized: $\Sigma_{\tilde{X}} = I_X$ and $\Sigma_{\tilde{Y}} = I_Y$*
3. *The $(i,j)^{th}$ element of the matrix $\Sigma_{\tilde{X}\tilde{Y}}$ is $\sqrt{1-\lambda_i}\delta_{ij}$. Thus X_i, Y_j have covariance $\sqrt{1-\lambda_i}$ if $i = j$, and are uncorrelated otherwise.*

To illustrate the above properties, we give below the full covariance matrix between variables with dimensions $n_x = 3, n_y = 2$

$$\Sigma_{\tilde{X}\tilde{Y},\tilde{X}\tilde{Y}} = \begin{pmatrix} 1 & 0 & 0 & \sqrt{1-\lambda_1} & 0 \\ 0 & 1 & 0 & 0 & \sqrt{1-\lambda_2} \\ 0 & 0 & 1 & 0 & 0 \\ \sqrt{1-\lambda_1} & 0 & 0 & 1 & 0 \\ 0 & \sqrt{1-\lambda_2} & 0 & 0 & 1 \end{pmatrix}$$

As a consequence of the above properties, we have the following corollary

Corollary 1 *For all $1 \le i \le \min(n_x, n_y)$*

$$\tilde{Y}_i = \sqrt{1-\lambda_i}\tilde{X}_i + N_i \quad with \quad N_i \sim \mathcal{N}(0, \sqrt{\lambda_i}). \tag{5}$$

Thus $Y^{(d)} = DX^{(d)} + \xi$ where $\xi \sim \mathcal{N}(0, N)$ and D, N are diagonal with $\sqrt{1-\lambda_i}, \sqrt{\lambda_i}$ on their diagonal.

This follows from the fact that two Gaussians can be presented as a noisy linear transformation, and since $E(\tilde{X}_i\tilde{Y}_i) = \sqrt{1-\lambda_i}$ and $V(\tilde{X}_i) = V(\tilde{Y}_i) = 1$ the transformation between $\tilde{X}_i, \tilde{Y}_i$ must be the one given above.

Also, since $\Sigma_{\tilde{X}|\tilde{Y}} = \Sigma_{\tilde{X}} - \Sigma_{\tilde{X}\tilde{Y}}\Sigma_{\tilde{Y}}^{-1}\Sigma_{\tilde{Y}\tilde{X}}$ and the matrices $\Sigma_{\tilde{X}|\tilde{Y}}, \Sigma_{\tilde{X}}$ have the structure given above, we have that

Corollary 2 *$\Sigma_{\tilde{X}|\tilde{Y}}$ is diagonal. Thus $\tilde{X}$ are conditionally independent given $\tilde{Y}$*

Since A_x, B_y are non singular, they do not affect the mutual information between X, Y and T

$$I(\tilde{X};T) = I(X;T) \qquad I(\tilde{Y};T) = I(Y;T)$$

Our optimization problem depends on X, Y only through the information values $I(X;T), I(Y;T)$, and thus we can use $\tilde{X}, \tilde{Y}$ instead of X, Y. Furthermore, the variables X cannot provide any information about $\tilde{Y}^{(n_x+1, n_y)}$ (when $n_y > n_x$), and thus we can assume that $n_y \leq n_x$ by discarding the excess Y variables.

From this point on, we will refer to the variables X and Y as if they were given in their canonical representation in the first place.

5 Proof of Theorem 1

We start by formulating $R(I_x)$ as a minimization problem. Using the Markov relation $T \to X \to Y$, we have $I(T;Y) = I(T;X) - I(T;X|Y)$ and thus

$$R(I_x) = \max_{p(t|x): T\to X\to Y, I(T;X)=I_x} I(T;Y) = I_x - \min_{p(t|x): T\to X\to Y, I(T;X)=I_x} I(T;X|Y) \quad .(6)$$

This, together with the form of $R^G(I_x)$ given in Equation 4 implies that Theorem 1 is equivalent to showing that

Theorem 2 *For every Markov chain $T \to X \to Y$, where $I(T;X) = I_x$ it holds that*

$$I(T;X|Y) \geq \frac{\hat{n}(I_x)}{2} \log \left(\prod_{i=1}^{\hat{n}(I_x)} (1-\lambda_i)^{\frac{1}{\hat{n}(I_x)}} + e^{\frac{2I_x}{\hat{n}(I_x)}} \prod_{i=1}^{\hat{n}(I_x)} \lambda_i^{\frac{1}{\hat{n}(I_x)}} \right) \tag{7}$$

The following lemma, which will be proven in section 6.1, bounds $I(T;X|Y)$ for subsets of the variable X.

Lemma 4 *For all d such that $d \leq n_x$, define $k = \min(n_y, d)$. Then*

$$I(T;X^{(d)}|Y) \geq \frac{k}{2} \log \left(\prod_{i=1}^{k} (1-\lambda_i)^{\frac{1}{k}} + e^{\frac{2I(T;X^{(d)})}{k}} \prod_{i=1}^{k} \lambda_i^{\frac{1}{k}} \right) \tag{8}$$

This of course provides a bound for $I(T;X|Y)$ by selecting $d = n_x$. However this bound is not tight, whereas that of Theorem 2 is.

We prove Theorem 2 by showing by induction that it holds for all values of $\hat{n}(I_x)$. The induction is over $k = n_y - \hat{n}(I_x)$ which is the number of segments between segment number $\hat{n}(I_x)$ and the last segment (note that we have at most n_y segments).

When $k = 0$, i.e. $\hat{n}(I_x) = n_y$ the claim follows from application of Lemma 4 with $d = n_x$.

In the induction step, we assume the theorem holds for I_x such that $n_y - \hat{n}(I_x) < k$ and show it holds for all I_x such that $n_y - \hat{n}(I_x) = k$. Thus from now on, we limit ourselves to I_x such that $\hat{n}(I_x) = n_y - k$. We separate X into two sets of variables. One contains its first $n_y - k$ elements $X^1 \equiv X^{(1,n_y-k)}$, and the other, $X^2 \equiv X^{(n_y-k+1,n_x)}$ contains the rest.

Note that

$$\begin{aligned} I(T;X|Y) &= I(T;X^1,X^2|Y) \\ &= h(X^1,X^2|Y) - h(X^1,X^2|T,Y) \\ &= h(X^1|Y) + h(X^2|Y) - h(X^1,X^2|T,Y) \\ &= h(X^1|Y) + h(X^2|Y) - h(X^1|T,Y) - h(X^2|X^1,T,Y) \\ &\geq h(X^1|Y) + h(X^2|Y) - h(X^1|T,Y) - h(X^2|T,Y) \\ &= I(T;X^1|Y) + I(T;X^2|Y) \end{aligned} \tag{9}$$

where the first equality is the definition of mutual information, the second equality is a result of X^1, X^2 being independent conditioned on Y, the third is the entropy chain rule, and the inequality is since conditioning reduces entropy.

Now, consider two variables T^1, T^2 such that the following holds:

$$\begin{aligned} I(T^1;X^1|Y) &= I(T;X^1|Y) \\ I(T^2;X^2|Y) &= I(T;X^2|Y) \\ I(T^1,T^2;X^1,X^2) &= I(T;X) \end{aligned} \tag{10}$$

and the Markov chains $T_1 \to X \to Y$, and $T_2 \to X \to Y$ are satisfied. Denote the set of variables which satisfy the above by $\mathcal{T}$ (note we are not placing any limitation on (T_1, T_2)). Then it immediately follows that $(T^1, T^2) = (T, T)$ is in $\mathcal{T}$. Therefor

$$\begin{aligned} I(T;X|Y) &\geq I(T;X^1|Y) + I(T;X^2|Y) \\ &\geq \min_{(T^1,T^2)\in\mathcal{T}} I(T^1;X^1|Y) + I(T^2;X^2|Y) \end{aligned} \tag{11}$$

where the first inequality follows from Equation 9 and the second from the fact that $\mathcal{T}$ contains (T,T).

Denote the eigenvalues corresponding to the variables in X^2 by $\lambda'_1, \ldots, \lambda'_k$.

To use the induction step, note that X^2 is $n_{X^2} = k$ dimensional and therefor $n_{X^2} - \hat{n}(I(T^2;X^2)) < k$. Thus we can bound $I(T^2;X^2|Y)$ using the induction step. Denote $n'_I = \hat{n}(I(T^2;X^2))$, then

$$I(T^2;X^2|Y) \geq \frac{n'_I}{2} \log \left(\prod_{i=1}^{n'_I} (1-\lambda'_i)^{\frac{1}{n'_I}} + e^{\frac{2I(T^2,X^2)}{n'_I}} \prod_{i=1}^{n'_I} {\lambda'}_i^{\frac{1}{n'_I}} \right) \tag{12}$$

We denote the function on the RHS by $f_2(I(T^2;X^2))$. We can also use Lemma 4 to bound $I(T^1;X^1|Y)$ so that

$$I(T^1;X^1|Y) \geq \frac{n_y-k}{2} \log \left(\prod_{i=1}^{n_y-k} (1-\lambda_i)^{\frac{1}{n_y-k}} + e^{\frac{2I(T^1,X^1)}{n_y-k}} \prod_{i=1}^{n_y-k} \lambda_i^{\frac{1}{n_y-k}} \right)$$

We denote the function on the RHS by $f_1(I(T^1;X^1))$. Thus we have

$$I(T^1;X^1|Y)+I(T^2;X^2|Y)\geq f_1(I(T^1;X^1))+f_2(I(T^2;X^2)) \tag{13}$$

The RHS of the above expression, which we want to minimize over $\mathcal{T}$ involves only $I(T^1;X^1),I(T^2;X^2)$. To minimize it over $\mathcal{T}$, we scan all values of $I(T^1;X^1),I(T^2;X^2)$ which are obtained in $\mathcal{T}$. To see what these values are, note that since X^1 and X^2 are independent, the following holds

$$\begin{aligned}
I(T;X)=I(T^1,T^2;X^1,X^2) &= h(X^1,X^2)-h(X^1,X^2|T^1,T^2)\\
&= h(X^1)+h(X^2)-h(X^1|T^1,T^2)-h(X^2|X^1,T^1,T^2)\\
&\geq h(X^1)+h(X^2)-h(X^1|T^1,T^2)-h(X^2|T^1,T^2)\\
&\geq h(X^1)+h(X^2)-h(X^1|T^1)-h(X^2|T^2)\\
&= I(T^1;X^1)+I(T^2;X^2)\quad .
\end{aligned}$$

The values of $I(T^1;X^1),I(T^2;X^2)$ which are obtained in $\mathcal{T}$ thus satisfy $I(T^1;X^1)+I(T^2;X^2)\leq I(T;X)$. This region is depicted in Figure 2.

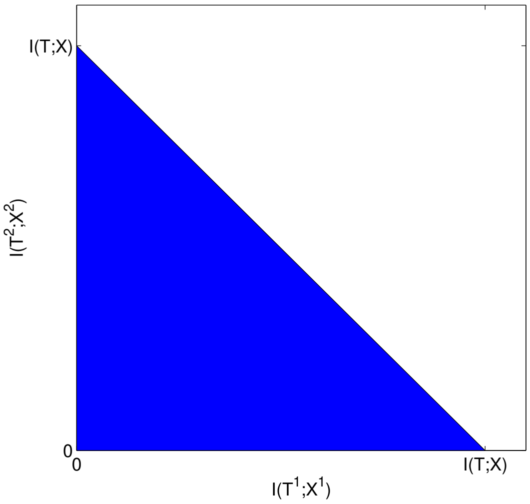

Figure 2: The shaded region shows the values of $I(T^1;X^1),I(T^2;X^2)$ which can be achieved in $\mathcal{T}$. Note that there may be points in the region which are not achievable, but all achievable points are included in it.

Thus the minimization in Equation 11 satisfies

$$\min_{(T^1,T^2)\in\mathcal{T}} f_1(I(T^1;X^1))+f_2(I(T^2;X^2))\geq \min_{0\leq\tilde{c}\leq I(T;X)}\ \min_{0\leq a\leq\tilde{c}} f_1(\tilde{c}-a)+f_2(a) \tag{14}$$

Where we have inequality since a given pair of values $I(T^1;X^1)=\tilde{c}-a, I(T^2;X^2)=a$ may not be achieved by an element in $\mathcal{T}$.

The minimum of Equation 14 is given in the following lemma, which is proven in section 6.2

Lemma 5 *If* $\hat{n}(I(T;X) = n_y - k$ *then*

$$\min_{0\le\tilde{c}\le I(T;X)} \min_{0\le a\le\tilde{c}} f_1(\tilde{c}-a) + f_2(a) = f_1(I(T;X)) \tag{15}$$

Thus we have that for $I(T;X)$ such that $\hat{n}(I(T;X)) = n_y - k$ the following holds (by combining Equations 11,14 and the definition of f_1)

$$I(T;X|Y) \ge \frac{n_y-k}{2}\log\left(\prod_{i=1}^{n_y-k}(1-\lambda_i)^{\frac{1}{n_y-k}} + e^{\frac{2I(T;X)}{n_y-k}}\prod_{i=1}^{n_y-k}\lambda_i^{\frac{1}{n_y-k}}\right) \tag{16}$$

which is what we wanted to prove.

6 Proof of Lemmas

In this section we give proofs for the Lemmas stated in the manuscript.

6.1 Proof of Lemma 4

In what follows we use $Y^{(d)} \equiv Y$ for $d > n_y$. First, we bound $I(T;X^{(d)}|Y)$ by an expression depending only on $Y^{(d)}$. We have that

$$I(T;X^{(d)}|Y) \ge I(T;X^{(d)}|Y^{(d)}) \tag{17}$$

which is a result of the following use of the information chain rule

$$\begin{aligned} I(X^{(d)};T,Y^{(d+1,n_y)}|Y^{(d)}) &= I(X^{(d)};Y^{(d+1,n_y)}|Y^{(d)}) + I(X^{(d)};T|Y) \\ &= I(X^{(d)};T|Y^{(d)}) + I(X^{(d)};Y^{(d+1,n_y)}|T,Y^{(d)}) \end{aligned}$$

Using the fact that $I(X^{(d)};Y^{(d+1,n_y)}|Y^{(d)}) = 0$ we get the desired inequality.

To simplify Equation 17 we use the following lemma:

Lemma 6 *The chain* $T \to X^{(d)} \to Y^{(d)}$ *is Markov*

Proof:
By the Markovity of $T \to X \to Y$ and using the information chain rule, we have

$$0 = I(Y;T|X) = I(Y^{(d)},Y^{(d+1,n_y)};T|X) \quad = \quad I(Y^{(d)};T|X) + I(Y^{(d+1,n_y)};T|X,Y^{(d)})$$

which implies

$$I(Y^{(d)};T|X) = 0 \tag{18}$$

Using the information chain rule

$$\begin{aligned} I(Y^{(d)};T,X^{(d+1,n_x)}|X^{(d)}) &= I(Y^{(d)};T|X^{(d)}) + I(Y^{(d)};X^{(d+1,n_x)}|T,X^{(d)}) \\ &= I(Y^{(d)};X^{(d+1,n_x)}|X^{(d)}) + I(Y^{(d)};T|X^{(d+1,n_x)},X^{(d)}) = 0 \end{aligned}$$

where we have used $I(Y^{(d)}; X^{(d+1,n_x)}|X^{(d)}) = 0$ and Equation 18. We have that a sum of two informations is zero, and therefore

$$I(Y^{(d)}; T|X^{(d)}) = 0 \tag{19}$$

which implies the desired Markov chain. $\square$

We can now write Equation 17 using the above Markov chain

$$\begin{aligned} I(T; X^{(d)}|Y) \geq I(T; X^{(d)}|Y^{(d)}) &= I(T; X^{(d)}) - I(T; Y^d) \\ &= I(T; X^{(d)}) - h(Y^{(d)}) + h(Y^{(d)}|T) \end{aligned} \tag{20}$$

If $d > n_y$ we have

$$I(T; X^{(d)}|Y) \geq I(T; X^{(d)}) - h(Y) + h(Y|T) \tag{21}$$

Equations 20,21 involve $h(Y^{(k)}|T)$ where $k \equiv \min(n_y, d)$. We now bound this conditional entropy. Using Corollary 1, write $h(Y^{(k)}|T) = h(DX^{(k)} + N|T)$. To bound this entropy, we use the conditional form of the entropy power inequality (see e.g. [1]), which states that if U, V, W, T are continuous random variables, and $U = V + W$ are n dimensional, then

$$e^{\frac{2}{n}h(U|T)} \geq e^{\frac{2}{n}h(V|T)} + e^{\frac{2}{n}h(W|T)} \tag{22}$$

Corollary 1 states that $Y^{(k)} = DX^{(k)} + N$ and therefor

$$\begin{aligned} e^{\frac{2}{k}h(Y^{(k)}|T)} &\geq e^{\frac{2}{k}h(DX^{(k)}|T)} + e^{\frac{2}{k}h(N|T)} \\ &= e^{\frac{2}{k}h(X^{(k)}|T) + \frac{2}{k}\log|D|} + e^{\frac{2}{k}h(N)} \\ &= e^{\frac{2}{k}h(X^{(k)}|T) + \frac{2}{k}\log|D|} + e^{\frac{1}{k}\log(2\pi e)^k|N|} \\ &= e^{-\frac{2}{k}I(T;X^{(k)}) + \frac{2}{k}h(X^{(k)})}|D|^{\frac{2}{k}} + (2\pi e)|N|^{\frac{2}{k}} \\ &= (2\pi e)e^{-\frac{2}{k}I(T;X^{(k)})}|D|^{\frac{2}{k}} + (2\pi e)|N|^{\frac{2}{k}} \\ &= (2\pi e)e^{-\frac{2}{k}I(T;X^{(k)})}\prod_{i=1}^{k}(1-\lambda_i)^{\frac{1}{k}} + (2\pi e)\prod_{i=1}^{k}\lambda_i^{\frac{1}{k}} \quad . \end{aligned} \tag{23}$$

Taking the log, $\frac{2}{k}h(Y^{(k)}|T) \geq \log 2\pi e + \log\left(e^{-\frac{2}{k}I(X^{(k)};T)}\prod(1-\lambda_i)^{\frac{1}{d}} + \prod\lambda_i^{\frac{1}{d}}\right)$ and subtracting $\frac{2}{k}h(Y^{(k)}) = \log(2\pi e)$, we obtain

$$-\frac{2}{k}I(Y^{(k)}; T) \geq \log\left(e^{-\frac{2}{k}I(T;X^{(k)})}\prod_{i=1}^{k}(1-\lambda_i)^{\frac{1}{k}} + \prod_{i=1}^{k}\lambda_i^{\frac{1}{k}}\right) \tag{24}$$

Substituting Equation 24 into 20, we finally obtain

$$\begin{aligned} I(T; X^{(d)}|Y) &= \frac{k}{2}\left(\frac{2}{k}I(T; X^{(k)}) - \frac{2}{k}I(T; Y^{(k)})\right) \\ &\geq \frac{k}{2}\log\left(e^{-\frac{2}{k}I(T;X^{(k)})}\prod_{i=1}^{k}(1-\lambda_i)^{\frac{1}{k}} + \prod_{i=1}^{k}\lambda_i^{\frac{1}{k}}\right) + \frac{k}{2}e^{\log\frac{2}{k}I(T;X^{(k)})} \\ &= \frac{k}{2}\log\left(\prod_{i=1}^{k}(1-\lambda_i)^{\frac{1}{k}} + e^{\frac{2}{k}I(T;X^{(k)})}\prod_{i=1}^{k}\lambda_i^{\frac{1}{k}}\right). \end{aligned} \tag{25}$$

6.2 Proof of Lemma 5

We first wish to minimize the function

$$F(a) = f_1(\tilde{c} - a) + f_2(a) \tag{26}$$

w.r.t the variable a.

In what follows, we show that $F(a)$ is monotonously increasing and therefore has its minimum at $a = 0$. The derivative of $F(a)$ w.r.t. a is given by

$$\begin{aligned}
\frac{d}{da}F(a) &= \frac{e^{\frac{2a}{n'_I}}\prod_{i=1}^{n'_I}{\lambda'}_i^{\frac{1}{k}}}{\prod_{i=1}^{n'_I}(1-{\lambda'}_i)^{\frac{1}{n'_I}} + e^{\frac{2a}{n'_I}}\prod_{i=1}^{n'_I}{\lambda'}_i^{\frac{1}{k}}} - \frac{e^{\frac{2(\tilde{c}-a)}{n_y-k}}\prod_{i=1}^{n_y-k}\lambda_i^{\frac{1}{n_y-k}}}{\prod_{i=1}^{n_y-k}(1-\lambda_i)^{\frac{1}{n_y-k}} + e^{\frac{2(\tilde{c}-a)}{n_y-k}}\prod_{i=1}^{n_y-k}\lambda_i^{\frac{1}{n_y-k}}} \\
&= \frac{1}{e^{-\frac{2a}{n'_I}}\prod_{i=1}^{n'_I}\left(\frac{1-{\lambda'}_i}{{\lambda'}_i}\right)^{\frac{1}{n'_I}}+1} - \frac{1}{e^{-\frac{2(\tilde{c}-a)}{n_y-k}}\prod_{i=1}^{n_y-k}\left(\frac{1-\lambda_i}{\lambda_i}\right)^{\frac{1}{n_y-k}}+1} \\
&= e^{-\frac{2(\tilde{c}-a)}{n_y-k}}\prod_{i=1}^{n_y-k}\left(\frac{1-\lambda_i}{\lambda_i}\right)^{\frac{1}{n_y-k}} - e^{-\frac{2a}{n'_I}}\prod_{i=1}^{n'_I}\left(\frac{1-{\lambda'}_i}{{\lambda'}_i}\right)^{\frac{1}{n'_I}}
\end{aligned}$$

where the last equality is up to a multiplicative positive constant.

Showing that $\frac{d}{da}F(a) > 0$ is equivalent to showing

$$\frac{e^{-\frac{2(\tilde{c}-a)}{n_y-k}}\prod_{i=1}^{n_y-k}\left(\frac{1-\lambda_i}{\lambda_i}\right)^{\frac{1}{n_y-k}}}{e^{-\frac{2a}{n'_I}}\prod_{i=1}^{n'_I}\left(\frac{1-{\lambda'}_i}{{\lambda'}_i}\right)^{\frac{1}{n'_I}}} \geq 1 \tag{27}$$

We develop this inequality and show that it must be true. Isolating a:

$$\begin{aligned}
e^{2a(\frac{1}{n_y-k}+\frac{1}{n'_I})} &\geq e^{\frac{2\tilde{c}}{n_y-k}}\prod_{i=1}^{n'_I}\left(\frac{1-{\lambda'}_i}{{\lambda'}_i}\right)^{\frac{1}{n'_I}}\prod_{i=1}^{n_y-k}\left(\frac{\lambda_i}{1-\lambda_i}\right)^{\frac{1}{n_y-k}} \qquad (28) \\
2a\frac{n_y-k+n'_I}{n'_I(n_y-k)} &\geq \frac{2\tilde{c}}{n_y-k} + \frac{1}{n'_I}\sum_{i=1}^{n'_I}\log\left(\frac{1-{\lambda'}_i}{{\lambda'}_i}\right) + \frac{1}{n_y-k}\sum_{i=1}^{n_y-k}\log\left(\frac{\lambda_i}{1-\lambda_i}\right) \\
a &\geq \frac{\tilde{c}n'_I}{n_y-k+n'_I} + \frac{n_y-k}{2(n_y-k+n'_I)}\sum_{i=1}^{n'_I}\log\left(\frac{1-{\lambda'}_i}{{\lambda'}_i}\right) \\
&+ \frac{n'_I}{2(n_y-k+n'_I)}\sum_{i=1}^{n_y-k}\log\left(\frac{\lambda_i}{1-\lambda_i}\right) \\
&= \frac{\tilde{c}n'_I}{n_y-k+n'_I} + \frac{n_y-k}{n_y-k+n'_I}\frac{1}{2}\sum_{i=1}^{n'_I}\log\left(\frac{1-{\lambda'}_i}{{\lambda'}_i}\right)\left(\frac{\lambda_{n_y-k+1}}{1-\lambda_{n_y-k+1}}\right) \\
&+ \frac{n'_I}{n_y-k+n'_I}\frac{1}{2}\sum_{i=1}^{n_y-k}\log\left(\frac{\lambda_i}{1-\lambda_i}\right)\left(\frac{1-\lambda_{n_y-k+1}}{\lambda_{n_y-k+1}}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\tilde{c}n'_I}{n_y - k + n'_I} + \frac{n_y - k}{n_y - k + n'_I}\frac{1}{2}\sum_{i=1}^{n'_I} \log\left(\frac{1-\lambda'_i}{\lambda'_i}\right)\left(\frac{\lambda_{n_y-k+1}}{1-\lambda_{n_y-k+1}}\right) \\
&- \frac{n'_I}{n_y - k + n'_I}\hat{c}(n_y - k + 1) \qquad (29)
\end{aligned}
$$

Note that the second term above is negative since λ_{n_y-k+1} is smaller than all the other λ in the sum, and $x/(1-x)$ is increasing. So every item in the log is smaller than 1. Thus, if we satisfy the inequality without the second term, we will surely satisfy it with the second term. The inequality thus becomes

$$
\begin{aligned}
a &\geq \frac{\tilde{c}n'_I}{n_y - k + n'_I} - \frac{n'_I}{n_y - k + n'_I}\hat{c}(n_y - k + 1) \qquad (30) \\
&= \frac{n'_I}{n_y - k + n'_I}\left(\tilde{c} - \hat{c}(n_y - k + 1)\right) \qquad (31)
\end{aligned}
$$

The last expression is negative, since we assumed that $\hat{n}(I(T;X)) = n_y - k$ and thus $\tilde{c} \leq I(T;X) < \hat{c}(n_y - k + 1)$. Since $a \geq 0$ by definition, the last inequality is always satisfied, which proves that $F(a)$ is increasing.

The minimum of $F(a)$ is thus achieved at $a = 0$, and its value is $F(0) = f_1(\tilde{c})$ since $f_2(0) = 0$. The overall minimization thus becomes

$$
\min_{0 \leq \tilde{c} \leq I(T;X)} \min_{0 \leq a \leq \tilde{c}} f_1(\tilde{c} - a) + f_2(a) = \min_{0 \leq \tilde{c} \leq I(T;X)} f_1(\tilde{c}) = f_1(I(T;X))
$$

where the last transition follows from $f_1(\tilde{c})$ being a decreasing function.

6.3 Proof of Lemma 3

We first observe that if v is a left eigenvector of $\Sigma_x^{-\frac{1}{2}}\Sigma_{x|y}\Sigma_x^{-\frac{1}{2}}$ with eigenvalue λ then $v\Sigma_x^{-\frac{1}{2}}$ is a left eigenvector of $\Sigma_{x|y}\Sigma_x^{-1}$ with the same eigenvalue. This can be seen from

$$
\begin{aligned}
v\Sigma_x^{-\frac{1}{2}}\Sigma_{x|y}\Sigma_x^{-\frac{1}{2}} &= \lambda v \\
\left(v\Sigma_x^{-\frac{1}{2}}\right)\Sigma_{x|y}\Sigma_x^{-1} &= \lambda v\Sigma_x^{-\frac{1}{2}}
\end{aligned}
$$

Since $\Sigma_x^{-\frac{1}{2}}\Sigma_{x|y}\Sigma_x^{-\frac{1}{2}}$ is a symmetric matrix, it has an orthonormal (and therefore independent) set of eigenvectors, which we denote by C_x. We have just shown that $A_x = C_x\Sigma_x^{-\frac{1}{2}}$ which implies that A_x is full rank, thus proving section 1 of the lemma. We similarly denote D_y the eigenvector matrix of $\Sigma_y^{-\frac{1}{2}}\Sigma_{y|x}\Sigma_y^{-\frac{1}{2}}$ and obtain a similar result for B_y.

To show that the covariance of $\tilde{X}$ is sphered (section 2 of the Lemma)

$$
\Sigma_{\tilde{X}} = A_x\Sigma_x A_x^T = C_x\Sigma_x^{-\frac{1}{2}}\Sigma_x\Sigma_x^{-\frac{1}{2}}C_x^T = C_xC_x^T = I
$$

and one can obtain $\Sigma_{\tilde{Y}} = B_y \Sigma_y B_y^T = I$ in a similar fashion.

To study the covariance between $\tilde{X}$ and $\tilde{Y}$ recall from Lemma 2 that B_y can be obtained from A_x via a transformation $B_y = RA_x \Sigma_{xy} \Sigma_y^{-1}$ where R is a scaling matrix, of size (n_y, n_x) with non-zero elements only on the diagonal. To find R we use the fact that $B_y \Sigma_y B_y^T = I$.

$$
\begin{aligned}
I = B_y \Sigma_y B_y^T &= RA_x \Sigma_{xy} \Sigma_y^{-1} \Sigma_y \Sigma_y^{-1} \Sigma_{yx} A_x^T R^T \\
&= RA_x \left(\Sigma_{xy} \Sigma_y^{-1} \Sigma_{yx} \Sigma_x^{-1} \right) \Sigma_x A_x^T R^T \\
&= R(I - D) A_x \Sigma_x A_x^T R \\
&= R(I - D) R^T
\end{aligned}
$$

Thus $R_{ii} = (1 - \lambda_i)^{-\frac{1}{2}}$. We can now find the covariance $\Sigma_{\tilde{X}\tilde{Y}}$

$$
\begin{aligned}
\Sigma_{\tilde{X}\tilde{Y}} &= A_x \Sigma_{xy} B_y^T = A_x \Sigma_{xy} \Sigma_y^{-1} \Sigma_{yx} A_x^T R^T \\
&= A_x \left(\Sigma_{xy} \Sigma_y^{-1} \Sigma_{yx} \Sigma_x^{-1} \right) \Sigma_x A_x^T R^T \\
&= (I - D) A_x \Sigma_x A_x^T R^T \\
&= (I - D) A_x \Sigma_x A_x^T R^T \\
&= (I - D) R^T
\end{aligned}
$$

which implies that the $(i,j)^t h$ element of $\Sigma_{\tilde{X}\tilde{Y}}$ is $\sqrt{1 - \lambda_i} \delta_{ij}$, as in Lemma 3.

Acknowledgments

We thank Gal Chechik and Eyal Krupka for helpful discussions and ideas regarding several parts of the proof. This work is partly supported by a grant from the Israeli Academy of Science. A.G. was supported by the Eshkol foundation.

References

[1] T. Berger abd R. Zamir. A semi-continuous version of the berger-yeung problem. *IEEE Transactions on Information Theory*, pages 1520–1526, 1999.

[2] G. Chechik, A. Globerson, N.Tishby, and Y.Weiss. Gaussian information bottleneck. In *Advances in neural information processing systems (Vol. 16)*. MIT Press, 2003.

[3] J.R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley and Sons, 1988.