THE DUAL INFORMATION BOTTLENECK

Anonymous authors
Paper under double-blind review

ABSTRACT

The Information Bottleneck (IB) framework is a general characterization of optimal representations in large scale learning that suggests an intriguing interpretation of efficient representations in deep learning in particular. It is based on the optimal trade off between the representation complexity and accuracy, both of which are quantified by mutual information. The problem is solved by alternating projections between the *encoder* and *decoder* of the representation, which can be performed locally at each representation level. The non-parametric IB framework, however, has several drawbacks: (i) the difficult estimation of information measures in high dimensions; (ii) the inability to extend the encoder beyond the training data - the need to to know the full joint distribution of the inputs and labels; (iii) the IB does not necessarily optimize the actual prediction of unseen labels, which can be critical when training the encoder from finite samples. Here we present a new framework, the Dual Information Bottleneck (dualIB), which improves all the above drawbacks. By switching the order in the KL-divergence between the representation decoder and data, formally the dual distortion, the optimal decoder becomes the geometric rather than the arithmetic mean of the input points. This conversion preserves exponential forms of the original distribution and consequently conserves the original features of the data. We show how to solve this new representation learning formulation and illustrate it's favorable properties over the original IB when trained from a small sample. We also analyze the critical points of the dualIB and discuss their importance for the quality of this approach.

1 INTRODUCTION

1.1 THE INFORMATION BOTTLENECK METHOD

The Information Bottleneck (IB) method (Tishby et al., 1999), is an information-theoretic framework for describing efficient representations of a large scale "input" random variable X (input patterns), for predicting an "output" variable Y (desired label). In this setting the joint distribution of X and Y, $p(\mathrm{x},\mathrm{y})$ defines the problem, or rule, and the training data are a finite sample from this distribution. In general, we assume that $p(\mathrm{y}\mid\mathrm{x})$ is strictly stochastic, and hence bounded away from $\{0,1\}$[1]. The representation variable $\hat{\mathrm{X}}$ is in general a stochastic function of X which forms a Markov chain $\mathrm{Y}\rightarrow\mathrm{X}\rightarrow\hat{\mathrm{X}}$, and only depends on Y through the input X. We call the map $p(\hat{\mathrm{x}}\mid\mathrm{x})$ the *encoder* of the representation and denote by $p(\mathrm{y}\mid\hat{\mathrm{x}})$ the *Bayes optimal decoder* for this representation; i.e., the best possible prediction of the *desired label* Y from the representation $\hat{\mathrm{X}}$.

The IB has direct successful applications for representation learning in various domains, from vision and speech processing (Ma et al., 2019), to neuroscience (Schneidman et al., 2001), and Natural Language Processing (Li & Eisner, 2019). Due to the notorious difficulty in estimating mutual information in high dimension, variational approximations to the IB have been suggested and applied also to Deep Learning (Alemi et al., 2016; Parbhoo et al., 2018).

For large scale learning, when almost all input patterns can be considered *typical* in the information theoretic sense (Cover & Thomas, 2006), the cardinality of the typical inputs is characterized by the pattern's entropy, $H(\mathrm{X})=-\sum_{\mathrm{x}}p(\mathrm{x})\log p(\mathrm{x})$. Under similar typicality assumption, the mutual

[1] This may seem like a limitation of the method, but it can be shown that in the deterministic limit the IB is well defined. This stochastic assumption simplifies the analysis but does not pose any real limitation.

information (MI) of the encoder, $I(\mathrm{X};\hat{\mathrm{X}}) = \sum_{\mathrm{x},\hat{\mathrm{x}}} p(\mathrm{x},\hat{\mathrm{x}}) \log \frac{p(\mathrm{x}|\hat{\mathrm{x}})}{p(\mathrm{x})}$ characterizes the cardinality of the typical representations, $\hat{\mathrm{X}}$. On the other hand, the mutual information of the Bayes optimal decoder, $I(\mathrm{Y};\hat{\mathrm{X}}) = \sum_{\mathrm{y},\hat{\mathrm{x}}} p(\mathrm{y},\hat{\mathrm{x}}) \log \frac{p(\mathrm{y}|\hat{\mathrm{x}})}{p(\mathrm{y})}$, is equivalent to the cross-entropy generalization error.

The IB trade off between the encoder and decoder mutual information values is defined by the minimization of the Lagrangian:

$$\mathcal{F}[p_\beta(\hat{\mathrm{x}} \mid \mathrm{x}); p_\beta(\hat{\mathrm{x}}); p_\beta(\mathrm{y} \mid \hat{\mathrm{x}})] = I(\mathrm{X};\hat{\mathrm{X}}) - \beta I(\mathrm{Y};\hat{\mathrm{X}}) \;, \tag{1}$$

independently over the convex sets of the normalized distributions, $\{p_\beta(\hat{\mathrm{x}} \mid \mathrm{x})\}$, $\{p_\beta(\hat{\mathrm{x}})\}$ and $\{p_\beta(\mathrm{y} \mid \hat{\mathrm{x}})\}$, given a positive Lagrange multiplier β. As shown in (Tishby et al., 1999; Shamir et al., 2010), this is a natural generalization of the classical concept of *Minimal Sufficient Statistics* (Cover & Thomas, 2006), where the estimated parameter is replaced by the output variable Y and *exact* statistical sufficiency is characterized by the mutual information equality: $I(\hat{\mathrm{X}};\mathrm{Y}) = I(\mathrm{X};\mathrm{Y})$. The minimality of the statistics is captured by the minimization of $I(\mathrm{X};\hat{\mathrm{X}})$, due to the Data Processing Inequality (DPI). However, non-trivial minimal sufficient statistics only exist for very special parametric distributions known as exponential families (Brown, 1986). Thus in general, the IB relaxes the minimal sufficiency problem to a continuous family of representations $\hat{\mathrm{X}}$ which are characterized by the trade off between compression, $I(\mathrm{X};\hat{\mathrm{X}}) \equiv I_{\mathrm{X}}$, and accuracy, $I(\mathrm{Y};\hat{\mathrm{X}}) \equiv I_{\mathrm{Y}}$, along a convex line in the *Information-Plane* (I_{Y} vs. I_{X}). When the rule $p(\mathrm{x},\mathrm{y})$ is strictly stochastic, the convex optimal line is smooth and each point along the line is uniquely characterized by the value of β. We can then consider the optimal representations $\hat{\mathrm{x}} = \hat{\mathrm{x}}(\beta)$ as encoder-decoder pairs: $(p_\beta(\mathrm{x}|\hat{\mathrm{x}}), p_\beta(\mathrm{y}|\hat{\mathrm{x}}))^2$ - a point in the continuous manifold defined by the Cartesian product of these distribution simplexes. We also consider a small variation of these representations, $\delta\hat{\mathrm{x}}$, as an infinitesimal change in this (encoder-decoder) continuous manifold (not necessarily on the optimal line(s)).

1.2 IB and Rate-Distortion Theory

The IB optimization trade off can be considered as a generalized rate-distortion problem (Cover & Thomas, 2006) with the distortion function between a data point, x and a representation point $\hat{\mathrm{x}}$ taken as the KL-divergence between their predictions of the desired label y:

$$d_{\mathrm{IB}}(\mathrm{x},\hat{\mathrm{x}}) = D[p(\mathrm{y} \mid \mathrm{x}) || p_\beta(\mathrm{y}|\hat{\mathrm{x}})] = \sum_{\mathrm{y}} p(\mathrm{y} \mid \mathrm{x}) \log \frac{p(\mathrm{y} \mid \mathrm{x})}{p_\beta(\mathrm{y} \mid \hat{\mathrm{x}})} \;. \tag{2}$$

The expected distortion $\mathbb{E}_{p_\beta(\mathrm{x},\hat{\mathrm{x}})}[d_{\mathrm{IB}}(\mathrm{x},\hat{\mathrm{x}})]$ for the optimal decoder is simply the label-information loss: $I(\mathrm{X};\mathrm{Y}) - I(\hat{\mathrm{X}};\mathrm{Y})$, using the Markov chain condition. Thus minimizing the expected IB distortion is equivalent to maximizing $I(\hat{\mathrm{X}};\mathrm{Y})$, or minimizing equation 1. Minimizing this distortion is equivalent to minimizing the log-loss or the cross-entropy loss, as done in most deep learning applications, and it upper-bounds other loss functions such as the $\mathcal{L}_1$-loss (due to the Pinsker inequality, or (Painsky & Wornell, 2018)). The Pinsker inequality shows that both orders of the cross-entropy upper bound to the $\mathcal{L}_1$-loss, or total-variation distance, $\min\{D[q||p], D[p||q]\} \geq \frac{1}{2\log 2}||p-q||_1^2$.

1.3 The IB Equations

For discrete X and Y, a necessary condition for the IB (local) minimization is given by the three self-consistent equations for the optimal encoder-decoder pairs, known as the IB *equations*:

$$\begin{cases} (i) & p_\beta(\hat{\mathrm{x}} \mid \mathrm{x}) = \frac{p_\beta(\hat{\mathrm{x}})}{Z(\mathrm{x};\beta)} e^{-\beta D[p(\mathrm{y}|\mathrm{x})||p_\beta(\mathrm{y}|\hat{\mathrm{x}})]} \\ (ii) & p_\beta(\hat{\mathrm{x}}) = \sum_{\mathrm{x}} p_\beta(\hat{\mathrm{x}} \mid \mathrm{x}) p(\mathrm{x}) \\ (iii) & p_\beta(\mathrm{y} \mid \hat{\mathrm{x}}) = \sum_{\mathrm{x}} p(\mathrm{y} \mid \mathrm{x}) p_\beta(\mathrm{x} \mid \hat{\mathrm{x}}) \end{cases} \;, \tag{3}$$

where $Z(\mathrm{x};\beta)$ is the normalization function. Iterating these equations is a generalized, Blahut-Arimoto, alternating projection algorithm (Tusnady & Csiszar, 1984; Cover & Thomas, 2006) and it converges to a stationary point of the Lagrangian, equation 1 (Tishby et al., 1999). Notice that the minimizing decoder, (equation 3-(iii)), is precisely the *Bayes optimal decoder* for the representation $\hat{\mathrm{x}}(\beta)$, given the Markov chain conditions.

[2] Here we use the *inverse encoder*, which is in the fixed dimension simplex of distributions over X.

1.4 Critical points and critical slowing down

One of the most interesting aspects of the IB equations is the existence of critical points along the optimal line of solutions in the *Information-Plane* (i.e. the information curve). At these points the representations change topology and cardinality (number of clusters) (Zaslavsky & Tishby, 2019; Parker et al., 2003) and they form the skeleton of the information curve and representation space. Under the strict stochastic assumption, the information-curve is a smooth function of the Lagrange multiplier β, but its derivative may not be smooth. Critical points are bifurcations of the solutions, which are values of β for which two different solutions (representations) co-exist. To identify such points we perform a perturbation analysis of the IB equations, as in (Zaslavsky & Tishby, 2019). Taking a small perturbation of the representation, denoted for brevity by $\delta\hat{\mathrm{x}}$, the changes in the $\log$ encoder and $\log$ decoder that satisfy equation 3 for a given β can be determined through the nonlinear eigenvalues problems[3]:

$$\left[I - \beta C^{\mathrm{IB}}_{\mathrm{xx}'}(\hat{\mathrm{x}}, \beta)\right] \frac{\partial \log p_\beta(\mathrm{x}' \mid \hat{\mathrm{x}})}{\partial \hat{\mathrm{x}}} = 0\ , \quad \left[I - \beta C^{\mathrm{IB}}_{\mathrm{yy}'}(\hat{\mathrm{x}}, \beta)\right] \frac{\partial \log p_\beta(\mathrm{y}' \mid \hat{\mathrm{x}})}{\partial \hat{\mathrm{x}}} = 0, \tag{4}$$

with the two square matrices C defined by:

$$C^{\mathrm{IB}}_{\mathrm{xx}'}(\hat{\mathrm{x}}, \beta) = \sum_{\mathrm{y}} p(\mathrm{y} \mid \mathrm{x}) \frac{p_\beta(\mathrm{x}' \mid \hat{\mathrm{x}})}{p_\beta(\mathrm{y} \mid \hat{\mathrm{x}})} p(\mathrm{y} \mid \mathrm{x}')\ , \quad C^{\mathrm{IB}}_{\mathrm{yy}'}(\hat{\mathrm{x}}, \beta) = \sum_{\mathrm{x}} p(\mathrm{y} \mid \mathrm{x}) \frac{p_\beta(\mathrm{x} \mid \hat{\mathrm{x}})}{p_\beta(\mathrm{y} \mid \hat{\mathrm{x}})} p(\mathrm{y}' \mid \mathrm{x}). \tag{5}$$

As shown in (Zaslavsky & Tishby, 2019), these two matrices have the same eigenvalues and have non-trivial eigenvectors (i.e., different co-existing optimal representations) precisely at the critical values of β, the bifurcation points of the IB solution. At these points the cardinality of the representation $\hat{\mathrm{X}}$ (the number of "IB-clusters") changes due to splits of clusters, resulting in topological phase transitions in the encoder. These critical points form the "skeleton" of the topology of the optimal representations. Between critical points the optimal representations change continuously (with β).

The important computational consequence of critical points is known as *critical slowing down* (Tredicce et al., 2004). For binary Y, near a critical point the convergence time, τ_β, of the iterations of equation 3 scales like: $\tau_\beta \sim 1/(1 - \beta\lambda_2)$, where λ_2 is the second eigenvalue of either $C^{\mathrm{IB}}_{\mathrm{yy}'}$ or $C^{\mathrm{IB}}_{\mathrm{xx}'}$. At criticality, $\lambda_2(\hat{\mathrm{x}}) = \beta^{-1}$ and the number of iterations diverges. This phenomenon dominates any local minimization of equation 3 which is based on alternate encoder-decoder optimization.

To illustrate the importance of the critical points and the critical slowing-down we consider a low-dimensional problem that is easy to visualize. Taking the joint distribution, $p(\mathrm{y}, \mathrm{x})$ with a binary Y and only 5 possible inputs. For a given β, the Blahut-Arimoto iterations converge to a stationary solutions of the IB *equations*, which amounts to discrete points in the interval $[0, 1]$ - the simplex of the $p(\mathrm{y})$ distributions - for each value of the Lagrange multiplier, β.

The evolution of the optimal decoder, $p_\beta(\mathbf{y} = 0 \mid \hat{\mathrm{x}})$, $\forall \hat{\mathrm{x}} \in \hat{\mathrm{X}}$, as a function of β, is a *bifurcation diagram* (Figure $1(a)$). The critical points of the IB equations correspond to bifurcations in the decoders' graph. At these critical points the non-trivial eigenvalues of $C^{\mathrm{IB}}_{\mathrm{yy}'}(\beta)$ cross the β^{-1} value (Figure $1(b)$), and the number of iterations diverges (critical slowing down) (Figure $1(c)$).

1.5 Drawbacks of the IB

The main theoretical and practical drawback of the IB is it's apparent reliance on the explicit knowledge of the joint distribution $p(\mathrm{x}, \mathrm{y})$. In the context of representation learning this is often considered a major flaw, as we are always given only a sample from the distribution. While the generalization properties of the finite sample IB representations have been studied in the past (Shamir et al., 2010), the main problem remains the extension of the *encoder*, $p(\hat{\mathrm{x}} \mid \mathrm{x})$, when trained from a small sample, to new unobserved patterns $\mathrm{x} \in \mathrm{X}$. The standard approach in representation learning is to assume (at least implicitly) the existence of features of x with their topology, or a specific parametric form for the encoder. The original IB is completely non-parametric and does not assume any topology on the input space, beyond label similarity. A related drawback of the IB is it's poor scalability with the size of the pattern-space X, which makes the reliable estimation of the encoder mutual information impractical for large learning problems.

[3] We ignore here the interactions between the representations, for simplicity.

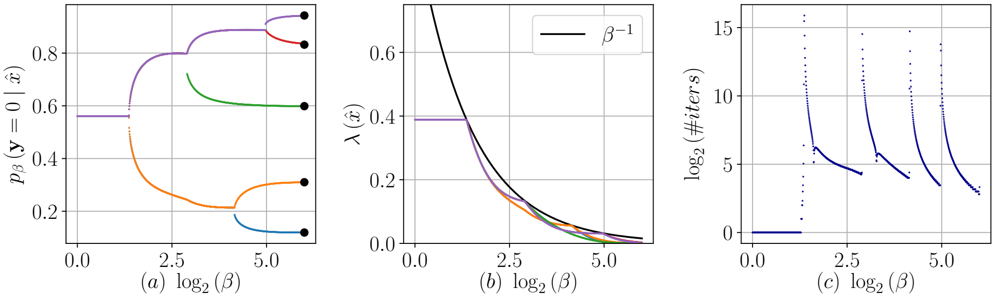

Figure 1: Example of the evolution of the IB decoder with growing β. (a) The *bifurcation diagram*; each color corresponds to one component of the representation $\hat{\mathrm{x}} \in \hat{\mathrm{X}}$ and depicts the decoder $p_\beta(\mathbf{y}=0 \mid \hat{\mathrm{x}})$, in the simplex $[0,1]$. The black dots correspond to the input distribution $p(\mathbf{y}=0 \mid \mathrm{x})$. (b) The second eigenvalue of $C_{\mathrm{yy'}}(\hat{\mathrm{x}};\beta)$, $\lambda_2(\hat{\mathrm{x}})$, along with β^{-1} as a function of β. (c) Convergence time of the algorithm as a function of β.

1.6 Contributions of this work

Our main contribution is by providing a general formal framework resolving the encoder extension drawback of the IB. While the use of specific features and variational approximations to the IB are not new, here we take a different approach assuming that $p(\mathrm{y} \mid \mathrm{x})$ can be approximated by exponential families. This defines a new rate-distortion problem, the Dual Infromation Bottleneck (dualIB).

We show that the new dualIB formulation preserves the sufficient statistics of the original distribution for all values of the trade off parameter, β. A desired property upon predictions of a parametrized distribution. Furthermore, it obtains a significant improvement of the finite sample label prediction. We demonstrate numerically that the dualIB encoder-decoder pairs improve the prediction of unseen labels when trained from small samples, compared to the original IB. We also discuss in detail the critical points of the new formulation and their relation to those of the original IB. Interestingly, as we elaborate in the following sections, all these desirable properties are achieved by merely switching the order of the arguments in the KL-divergence of the IB distortion.

The remainder of the paper is organized as follows. We present and solve the dualIB problem in §2. We discuss the dualIB's critical points in §2.1 and prove a general relation between the critical points of the dual and original IB in §2.2. In particular, we show that the dual critical points give the best approximation to the original IB information curve. In §3 we discuss the dualIB for exponential families, dualExpIB, and demonstrate the advantages of our framework in this case. We also demonstrate numerically the improvement in the label prediction of the dualIB when trained from a small sample. We conclude in §4 with further extensions and possible applications.

2 The dualIB: Maximizing the Prediction Information

Supervised learning is generally separated into two phases: the training phase, where the internal representations are formed from the training data, and the prediction phase, where these representations are used to predict labels of new input patterns. We add to our Markov chain another variable, $\hat{\mathrm{Y}}$, the *predicted label* which obtains the same values as Y, but distributed differently:

$$\rlap{$\overbrace{\phantom{\mathrm{Y} \to \mathrm{X} \to \hat{\mathrm{X}}_\beta}}^{\text{training}}$}\mathrm{Y} \to \underbrace{\mathrm{X} \to \hat{\mathrm{X}}_\beta \to \hat{\mathrm{Y}}}_{\text{prediction}} . \tag{6}$$

The left-hand part of this chain describes the representation training, and the right-hand part is the Maximum Likelihood (ML) prediction using these representations (Slonim et al., 2006). So far the prediction variable $\hat{\mathrm{Y}}$ has not been part of the IB optimization problem. It has been implicitly assumed that the *Bayes optimal decoder*, $p_\beta(\mathrm{y} \mid \hat{\mathrm{x}})$, which minimizes the IB distortion at a given β, is also the best choice for making predictions of $\hat{\mathrm{Y}}$ from the representation $\hat{\mathrm{X}}_\beta$ through the right-hand Markov chain. That is, denoting $p_\beta(\hat{\mathrm{y}} \mid \hat{\mathrm{x}}) \equiv p_\beta(\mathrm{y} \mid \hat{\mathrm{x}})$ the ML decoder is the mixture over the internal representations:

$$p_\beta(\hat{\mathrm{y}} \mid \mathrm{x}) \equiv \sum_{\hat{\mathrm{x}}} p_\beta(\hat{\mathrm{y}} \mid \hat{\mathrm{x}}) p_\beta(\hat{\mathrm{x}} \mid \mathrm{x}). \tag{7}$$

This, however, is not necessarily optimal when the encoder and decoder are trained from finite samples (Shamir et al., 2010), or when the decoder is not Bayes optimal. We define the dualIB distortion by merely switching the order of the arguments in the KL-divergence of the original IB distortion, namely:

$$d_{\mathrm{dualIB}}(\mathrm{x}, \hat{\mathrm{x}}) = D[p_\beta(\mathrm{y} \mid \hat{\mathrm{x}}) || p(\mathrm{y} \mid \mathrm{x})] = \sum_{\mathrm{y}} p_\beta(\mathrm{y} \mid \hat{\mathrm{x}}) \log \frac{p_\beta(\mathrm{y} \mid \hat{\mathrm{x}})}{p(\mathrm{y} \mid \mathrm{x})} \ , \tag{8}$$

which in geometric terms is known as the *dual* distortion problem (Felice & Ay, 2019).

That is, given the “input” random variable X (input patterns), the ”output” random variable Y (desired label) and the “representation” $\hat{\mathrm{X}}$, the ML optimal “predicted label” $\hat{\mathrm{Y}}$ is given by the mixture distribution, equation 7.[4]

The dualIB optimization can be written as the following rate-distortion problem:

$$\mathcal{F}^*[p_\beta(\hat{\mathrm{x}} \mid \mathrm{x}); p_\beta(\hat{\mathrm{x}}); p_\beta(\mathrm{y} \mid \hat{\mathrm{x}})] = I(\mathrm{X}; \hat{\mathrm{X}}) + \beta \mathbb{E}_{p(\mathrm{x},\hat{\mathrm{x}})}[d_{\mathrm{dualIB}}(\mathrm{x}, \hat{\mathrm{x}})] \ , \tag{9}$$

with the average distortion given in terms of mutual information on $\hat{\mathrm{Y}}$, $I(\hat{\mathrm{X}}; \hat{\mathrm{Y}})$ and $I(\mathrm{X}; \hat{\mathrm{Y}})$:

$$\begin{aligned} \mathbb{E}_{p(\mathrm{x},\hat{\mathrm{x}})}[d_{\mathrm{dualIB}}(\mathrm{x}, \hat{\mathrm{x}})] &= \underbrace{I(\hat{\mathrm{X}}; \hat{\mathrm{Y}}) - I(\mathrm{X}; \hat{\mathrm{Y}})}_{(a)} + \underbrace{\mathbb{E}_{p(\mathrm{x})}[D[p_\beta(\hat{\mathrm{y}} \mid \mathrm{x}) \| p(\hat{\mathrm{y}} \mid \mathrm{x})]]}_{(b)} \\ &\geq I(\hat{\mathrm{X}}; \hat{\mathrm{Y}}) - I(\mathrm{X}; \hat{\mathrm{Y}}) \ . \end{aligned} \tag{10}$$

This is similar to the known IB relation: $\mathbb{E}_{p(\mathrm{x},\hat{\mathrm{x}})}[d_{\mathrm{IB}}(\mathrm{x}, \hat{\mathrm{x}})] = I(\mathrm{Y}; \mathrm{X}) - I(\mathrm{Y}; \hat{\mathrm{X}})$ with an extra positive term (b).

Both terms, (a) and (b), vanish precisely when $\hat{\mathrm{X}}$ is a sufficient statistic for X with respect to $\hat{\mathrm{Y}}$, since we can then reverse the order of X and $\hat{\mathrm{X}}$ in the Markov chain (equation 6). This replaces the roles of Y and $\hat{\mathrm{Y}}$ as the variable for which $\hat{\mathrm{X}}_\beta$ are approximately minimally sufficient statistics. In that sense the dualIB shifts the emphasis from the training phase to the prediction phase. This implies that minimizing the dualIB functional maximizes the mutual information between Y and $\hat{\mathrm{Y}}$, $I(\mathrm{Y}; \hat{\mathrm{Y}})$, as well as the mutual information $I(\mathrm{X}; \hat{\mathrm{Y}})$. This is illustrated in figure 4(b)-*c*.

The next theorem states the form of the solutions of *the Dual Information Bottleneck*:

Theorem 1. *The minima of equation 9, can be obtained by generalized Blahut-Arimoto iterations between the encoder and the decoder as in the original* IB*, with the following modifications: (i) Replace the distortion by its dual in the encoder update; (ii) Update the decoder by the encoder's ”geometric” mean of the data distributions* $p(\mathrm{y} \mid \mathrm{x})$.

The proof is given in §A.2.

The alternating projections between the encoder and decoder, which converge to a solution of the dualIB at a given value of the Lagrange multiplier β, are implemented by the following iterative algorithm[5]:

[4] We use the notation $p(\hat{\mathrm{y}} \mid \mathrm{x}) \equiv p(\mathrm{y} \mid \mathrm{x})$, when there is no β subscript.

[5] Unless stated otherwise, $\log = \ln \ \equiv \log_e$

Algorithm 1 dualIB iterative algorithm

```
1: while $\left|p_\beta^{t+1}(\mathrm{y} \mid \hat{\mathrm{x}}) - p_\beta^{t}(\mathrm{y} \mid \hat{\mathrm{x}})\right| > \epsilon$ do
2:     $Z_{\hat{\mathbf{x}}|\mathbf{x}}^{t+1}(\mathrm{x};\beta) = \sum_{\hat{\mathrm{x}}} p_\beta^t(\hat{\mathrm{x}}) e^{-\beta D\left[p_\beta^t(\mathrm{y}|\hat{\mathrm{x}}) \| p(\mathrm{y}|\mathrm{x})\right]}$
3:     $p_\beta^{t+1}(\hat{\mathrm{x}} \mid \mathrm{x}) = \frac{p_\beta^t(\hat{\mathrm{x}})}{Z_{\hat{\mathbf{x}}|\mathbf{x}}^{t+1}(\mathrm{x};\beta)} e^{-\beta D\left[p_\beta^t(\mathrm{y}|\hat{\mathrm{x}}) \| p(\mathrm{y}|\mathrm{x})\right]}$
4:     $p_\beta^{t+1}(\hat{\mathrm{x}}) = \sum_{\mathrm{x}} p_\beta^{t+1}(\hat{\mathrm{x}} \mid \mathrm{x}) p(\mathrm{x})$
5:     $Z_{\mathrm{y}|\hat{\mathrm{x}}}^{t+1}(\hat{\mathrm{x}};\beta) = \sum_{\mathrm{y}} e^{\sum_{\mathrm{x}} p_\beta^{t+1}(\mathrm{x}|\hat{\mathrm{x}}) \log p(\mathrm{y}|\mathrm{x})}$
6:     $p_\beta^{t+1}(\mathrm{y} \mid \hat{\mathrm{x}}) = \frac{1}{Z_{\mathbf{y}|\hat{\mathbf{x}}}(\hat{\mathrm{x}};\beta)} e^{\sum_{\mathrm{x}} p_\beta^{t+1}(\mathrm{x}|\hat{\mathrm{x}}) \log p(\mathrm{y}|\mathrm{x})}$
7: return $p_\beta(\hat{\mathrm{x}}), p_\beta(\hat{\mathrm{x}} \mid \mathrm{x}), p_\beta(\mathrm{y} \mid \hat{\mathrm{x}})$
```

Here the Z's are standard normalization factors (partition functions).
As in the IB, the encoder update (row 3) and decoder update (row 6) are the core of the algorithm.

Figure 2 illustrates the behavior of the dualIB algorithm for the same distribution as in Figure 1. The overall behavior is very similar, but the critical points (representation splits) slightly change their location. In the following sections we provide a detailed analysis and comparison between the two frameworks.

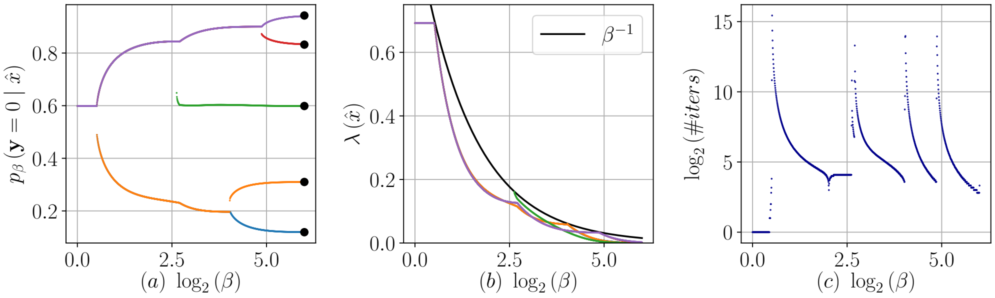

Figure 2: Same as Figure 1 for the dualIB solutions with the $C_{\mathrm{yy'}}^{\mathrm{dualIB}}$ matrix. (a) The representations bifurcate at the critical values of β. (b) Critical points appear when an eigenvalue (color) crosses the β^{-1} (black) line. (c) Near these points there is *critical slowing down* and the numbers of iterations of Algorithm 1 diverge.

2.1 The Critical Points of the dualIB

As discussed in §1.4 the skeleton of the IB optimal bound (the information curve) is constituted by the critical points in which the topology (cardinality) of the representation changes. A similar stability analysis of the dualIB equations reveals similar conditions for the critical points.

Theorem 2. *The* dualIB *critical points are detected by non-trivial solutions of the nonlinear eigenvalue problem:*

$$\left[I - \beta C_{\mathrm{xx'}}^{\mathrm{dualIB}}(\hat{\mathrm{x}},\beta)\right] \frac{\partial \log p_\beta(\mathrm{x'} \mid \hat{\mathrm{x}})}{\partial \hat{\mathrm{x}}} = 0\,, \quad \left[I - \beta C_{\mathrm{yy'}}^{\mathrm{dualIB}}(\hat{\mathrm{x}},\beta)\right] \frac{\partial \log p_\beta(\mathrm{y'} \mid \hat{\mathrm{x}})}{\partial \hat{\mathrm{x}}} = 0, \tag{11}$$

with the matrices C^{dualIB} given by:

$$C_{\mathrm{xx'}}^{\mathrm{dualIB}}(\hat{\mathrm{x}};\beta) = \sum_{\mathrm{y},\tilde{\mathrm{y}},\tilde{\mathrm{x}}} p_\beta(\mathrm{y} \mid \hat{\mathrm{x}}) p_\beta(\tilde{\mathrm{x}} \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y} \mid \mathrm{x})}{p(\mathrm{y} \mid \tilde{\mathrm{x}})} \cdot p_\beta(\mathrm{x'} \mid \hat{\mathrm{x}}) p_\beta(\tilde{\mathrm{y}} \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y} \mid \mathrm{x'})}{p(\tilde{\mathrm{y}} \mid \mathrm{x'})}$$

$$C_{\mathrm{yy'}}^{\mathrm{dualIB}}(\hat{\mathrm{x}};\beta) = \sum_{\mathrm{x},\tilde{\mathrm{x}},\tilde{\mathrm{y}}} p_\beta(\mathrm{x} \mid \hat{\mathrm{x}}) p_\beta(\tilde{\mathrm{y}} \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y} \mid \mathrm{x})}{p(\tilde{\mathrm{y}} \mid \mathrm{x})} \cdot p_\beta(\mathrm{y'} \mid \hat{\mathrm{x}}) p_\beta(\tilde{\mathrm{x}} \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y'} \mid \mathrm{x})}{p(\mathrm{y'} \mid \tilde{\mathrm{x}})}. \tag{12}$$

The proof to *theorem* 2 is given in §A.3.

Lemma 3. *The matrices $C^{\text{dualIB}}_{\mathrm{xx}'}(\hat{\mathrm{x}};\beta), C^{\text{dualIB}}_{\mathrm{yy}'}(\hat{\mathrm{x}};\beta)$ have the same eigenvalues $\{\lambda_i\}$, with $\lambda_1(\hat{\mathrm{x}}) = 0$. With binary* Y*, the critical points are obtained at $\lambda_2(\hat{\mathrm{x}}) = \beta^{-1}$.*

The proof of *lemma* 3 given in section §A.3.1.

As in the IB, at the critical points, β_c^{dualIB}, the partial derivatives of the encoder and decoder with respect to β, $\partial \log p_\beta(\mathrm{x} \mid \hat{\mathrm{x}})/\partial\beta$, $\partial \log p_\beta(\mathrm{y} \mid \hat{\mathrm{x}})/\partial\beta$, have multiple (at least two) values. This results in discontinuities (cusps) in the encoder and decoder mutual information values as functions of β along the optimal line, with an undefined second derivative.

2.2 The Information Plane of the dualIB

The *Information Plane*, $I_{\mathrm{X}} = I(\hat{\mathrm{X}};\mathrm{X})$ vs. $I_{\mathrm{Y}} = I(\hat{\mathrm{X}};\mathrm{Y})$, is the standard visualization of the compression-prediction trade off of the IB. It can be defined for *any* encoder once the decoder is the Bayes optimal (equation 3-(iii)), for which the I_{Y} is the actual information of the representation on the desired label (Tishby et al., 1999).

Comparing the dualIB information curve to the IB curve shows the quality of this approximation. Figure 3(a) depicts this comparison. While we know that $I_{\mathrm{Y}}^{\mathrm{IB}}(\beta)$ is always higher, the two curves are almost indistinguishable. To better understand the relationship between these two curves, we look at the values of I_{X} and I_{Y} as functions of the corresponding β (Figure 3(b),(c)). The important role of the critical points is revealed as the corresponding cusps along these curves. As we argue below, the IB information values are strictly below those of the dualIB, but the distance between them is minimized precisely at the dual critical points.

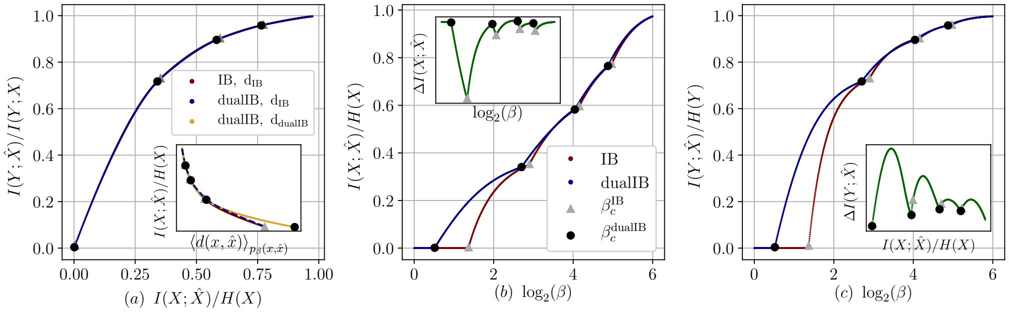

Figure 3: The IB's and dualIB's *Information Plane*. (a) I_{Y} vs. I_{X} for the two algorithms. The black dots are the dualIB critical points, β_c^{dualIB}, and the grey triangles are the IB critical points, β_c^{IB}. The corresponding distortion functions are shown in the inset. (b) The functions $I_{\mathrm{X}}^{\mathrm{IB}}(\beta)$ and $I_{\mathrm{X}}^{\text{dualIB}}(\beta)$. Both curves are monotonic and concave between the critical points. The inset indicates the relative difference between the curves, where the alternating order of the critical points is clearly observed. (c) Similarly, $I_{\mathrm{Y}}^{\mathrm{IB}}(\beta)$ and $I_{\mathrm{Y}}^{\text{dualIB}}(\beta)$ are monotonic and piece-wise concave. The relative discrepancy between the information curves is clearly minimized at the dualIB critical points (inset). The functions approach each other for large β.

Lemma 4. *$I_{\mathrm{X}}(\beta)$ and $I_{\mathrm{Y}}(\beta)$, along the optimal lines, are non-decreasing piece-wise concave, functions of β. When their second derivative (with respect to β) is defined, it is strictly negative.*

Lemma 5. *For any sub-optimal information curve ($I_{\mathrm{X}}, I_{\mathrm{Y}}$), $I_{\mathrm{X}}^{\mathrm{IB}}(\beta) \le I_{\mathrm{X}}(\beta)$ and $I_{\mathrm{Y}}^{\mathrm{IB}}(\beta) \le I_{\mathrm{Y}}(\beta)$, for all values of β.*

Proofs of the above are given in §A.4.

The information plane properties are summarized by the following theorem.

Theorem 6. *(i) The critical points of the two algorithms alternate: for each critical point, $\beta_c^{\text{dualIB}} \le \beta_c^{\mathrm{IB}}$. (ii) The distance between the two information curves is minimized precisely at the* dualIB *critical points β_c^{dualIB}. (iii) The two curves approach each other as $\beta \to \infty$.*

Proof. The proof follows from *lemmas* 4 and 5, together with the critical points analysis above, and is only sketched here. As the encoder and decoder at the critical points, β_c^{IB} and β_c^{dualIB}, have different left and right derivatives, they form cusps in the curves of the mutual information (I_{X} and I_{Y}) as functions of β. These cusps can only be consistent with the optimality of the IB curves if $\beta_c^{\mathrm{dualIB}} < \beta_c^{\mathrm{IB}}$ (this is true for any sub-optimal distortion), otherwise the curves intersect.

Moreover, at the dualIB critical points, the distance between the curves is minimized due to the strict concavity of the functions segments between the critical points. As the critical points imply discontinuity in the derivative, this results in a ”jump” in the information values. Therefore, at any $\beta_c^{\mathrm{dualIB}}$ the distance between the curves has a (local) minimum. This is depicted in Figure 3, comparing $I_{\mathrm{X}}(\beta)$ and $I_{\mathrm{Y}}(\beta)$ and their differences for the two algorithms.

The two curves approach each other for large β since the two distortion functions become close in the low distortion limit (as long as $p(\mathrm{y} \mid \mathrm{x})$ is bounded away from 0). □

3 The Exponential Family dualIB

As discussed in §1, one of the major drawbacks of the IB is the extension of the decoder inputs from an empirical sample to new unseen patterns. To address this issue we can assume the existence of features, or statistics of the pattern, $A_r(\mathrm{x})$, which are sufficient or approximately sufficient for the prediction of the label Y. That is, we consider the rule's distribution, $p(\mathrm{y} \mid \mathrm{x})$, to be an exponential family with these features. Exponential families form the elegant theoretical core of parametric statistics and often emerge as maximum entropy (Jaynes, 1957) or stochastic equilibrium distributions, subject to observed expectation constraints. They also form the class of parametric distributions for which exact, finite dimensional and additive, Minimal Sufficient Statistics exist (Kullback, 1959). Such distributions should also best match the rational behind the IB as extracting approximate sufficient statistics.

A key property of exponential families is their invariance under geometric averages. For this technical reason the dualIB encoder and decoder preserve the exponential form of the original data distribution for all values of β. Furthermore, the new (assumed limited to a finite dimension d) set of features/statistics are precisely the empirical expectations of these features. These d empirical averages are the only numbers needed for predicting the label at the required accuracy, hence their advantage in reducing the encoder's computational complexity.

Assuming that the rule distribution is of the form:

$$p(\mathrm{y} \mid \mathrm{x}) = e^{-\sum_{r=0}^{d} \lambda^r(y) A_r(x)} = e^{-\boldsymbol{\lambda}(\mathrm{y})\cdot \boldsymbol{A}(\mathrm{x}) - \log Z_{\mathbf{y}|\mathbf{x}}(\mathrm{x})} = \prod_{r=0}^{d} e^{-\lambda^r(\mathrm{y}) A_r(\mathrm{x})} , \tag{13}$$

where $A_r(\mathrm{x})$ are d linearly independent functions of the input x and $\lambda^r(\mathrm{y})$ are functions of the label y, or the parameters of this exponential family. The $\boldsymbol{\lambda}(\mathrm{y})$ can also be considered Lagrange multipliers associated with the constraints conditional expectations $\mathbb{E}_{p(\mathrm{x}|\mathrm{y})}[A_r(\mathrm{x})]$ in entropy maximization. The normalization factors, $Z_{\mathbf{y}|\mathbf{x}}(\mathrm{x})$, are written, for brevity, as $\lambda^0_{\mathbf{x}} \equiv \log(\sum_{\mathrm{y}} \prod_{r=1}^{d} e^{-\lambda^r(\mathrm{y}) A_r(\mathrm{x})})$ with $A_0(\mathrm{x}) \equiv 1$. We do not constrain the marginal $p(\mathrm{x})$.

The important fact about the exponential form is that all the mutual information, $I(\mathrm{X};\mathrm{Y})$, is fully captured by the d conditional expectations, $\mathbb{E}_{p(\mathrm{x}|\mathrm{y})}[A_r(\mathrm{x})]$, since these are the (minimal) sufficient statistics for the parameters. This means that all the relevant information (in the training sample) is captured solely by d-dimensional empirical expectations. This can lead to a huge reduction in computational complexity (from $\dim(\mathrm{X})$ to d).

We next show that for the dualIB, this dimension reduction is preserved or improved along the dual information curve, for all values of β.

Theorem 7. (dualExpIB*) For data from an exponential family, equation 13, the optimal decoder of the* dualIB*, at a given β, is given by:*

$$p_\beta(\mathrm{y} \mid \hat{\mathrm{x}}) = e^{-\sum_{r=1}^{d} \lambda^r(\mathrm{y}) A_{r,\beta}(\hat{\mathrm{x}}) - \lambda^0_\beta(\hat{\mathrm{x}})} , \quad \lambda^0_\beta(\hat{\mathrm{x}}) = \log(\sum_{\mathrm{y}} e^{-\sum_{r=1}^{d} \lambda^r(\mathrm{y}) A_{r,\beta}(\hat{\mathrm{x}})}) , \tag{14}$$

and the respective encoder:

$$p_\beta(\hat{\mathrm{x}} \mid \mathrm{x}) = \frac{p_\beta(\hat{\mathrm{x}}) e^{-\beta \lambda^0_\beta(\hat{\mathrm{x}})}}{Z_{\hat{\mathbf{x}}|\mathbf{x}}(\mathrm{x};\beta)} e^{-\beta \sum_{r=1}^{d} \lambda^r_\beta(\hat{\mathrm{x}})[A_r(\mathrm{x}) - A_{r,\beta}(\hat{\mathrm{x}})]} \ , \tag{15}$$

with the constraints and multipliers expectations,

$$A_{r,\beta}(\hat{\mathrm{x}}) \equiv \sum_{\mathrm{x}} p_\beta(\mathrm{x} \mid \hat{\mathrm{x}}) A_r(\mathrm{x}) \ , \ \lambda^r_\beta(\hat{\mathrm{x}}) \equiv \sum_{\mathrm{y}} p_\beta(\mathrm{y} \mid \hat{\mathrm{x}}) \lambda^r(\mathrm{y}) \ , \ 1 \le r \le d \ . \tag{16}$$

The complete derivations for this section are given in §A.5.

This defines a simplified iterative algorithm to solve the $\mathrm{dualExpIB}$ problem, since we can replace the decoders' update rule at each iteration in the dualIB algorithm (Algorithm 1, row 6) with the simplified expression given in equation 14.

As a concrete illustration to the performance of the dualExpIB we examine a problem in which the rule distribution is given by an exponential family, $p(\mathrm{y} \mid \mathrm{x}) \sim \mathcal{N}(0, a + b|\sin(c\pi \mathrm{x})|)$, as in Globerson & Tishby (2003). Setting the constants a, b, c we produce a sample S (of size $N = 1000$) on which we calculate the empirical distribution, $\hat{p}(\mathrm{y} \mid \mathrm{x})$ and the empirical sufficient statistics $\hat{A}_r(\mathrm{x})$. Given the assumption on the rule we assess the exponential form empirical distribution, $\hat{p}_{\exp}(\mathrm{y} \mid \mathrm{x}; \hat{A}_r(\mathrm{x}))$. We then evaluate the IB on both empirical distributions, $\hat{p}$ and $\hat{p}_{\exp}$, and compare it to the dualExpIB given $\hat{A}_r(\mathrm{x}), \lambda^r(\mathrm{y})$[6].

The comparison between the IB and the dualExpIB, is given in Figure 4(a) as function of the continuous $\hat{\mathrm{y}}$ for a fixed $\hat{\mathrm{x}}$. It can be seen that the dualExpIB better captures the structure of the original data and is much less affected by the sampling noise. Next, we consider the *Information Plane* with respect to $\hat{\mathrm{Y}}$, where the expected maximization by the dualExpIB is observed, Figure 4(b). To evaluate the prediction ability of the dualExpIB we consider the $\mathcal{L}_1$*-expected loss* with respect to the original distribution. Looking at Figure 4(c) it is apparent that for all values of the trade off parameter, β, the dualExpIB minimizes the loss providing the superiority in prediction.

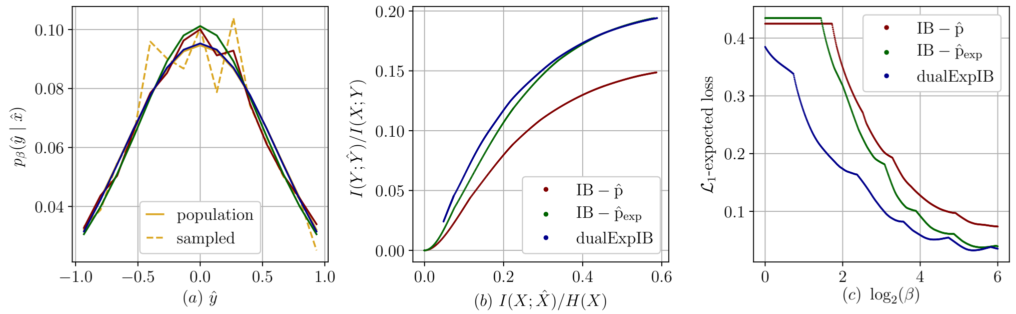

Figure 4: The IB's and dualExpIB's behavior over data sampled from $p(\mathrm{y} \mid \mathrm{x}) \sim \mathcal{N}(0, a + b|\sin(c\pi\mathrm{x})|)$ and $p(\mathrm{x}) \sim \mathcal{U}(-1, 1)$. (a) The decoder $p_\beta(\hat{\mathrm{y}} \mid \hat{\mathrm{x}})$ as a function of $\hat{\mathrm{y}}$ for some $\hat{x} \in (-1, 1)$ at $\log_2(\beta) = 4$. (b) The *Information Plane* with respect to $\hat{\mathrm{Y}}$, $I(\mathrm{Y}; \hat{\mathrm{Y}})$ vs. $I(\mathrm{X}; \hat{\mathrm{X}})$. (c) The $\mathcal{L}_1$*-expected loss*, $\mathbb{E}_{p(\mathrm{x})}[\|p(\hat{\mathrm{y}} \mid \mathrm{x}) - p_\beta(\hat{\mathrm{y}} \mid \mathrm{x})\|]$, as a function of the trade off parameter β.

4 Conclusion

We presented here the Dual Information Bottleneck (dualIB), a framework to preform the Information Bottleneck (IB) compression in a parametric setting. We have shown that the dualIB has several interesting properties: (i) it provides a good approximation to the original IB while using the low relevant dimensions of the original data. This can significantly reduce the complexity of finding good IB representations; (ii) it optimizes the information between the representation and the *predicted label* rather than the desired label as in the original IB. As we have shown, this can improve

[6] $\lambda^r(\mathrm{y})$ are defined by the specific exponential family.

the generalization error when trained on small samples since the predicted label is the one used in practice; (iii) It preserves the exponential form of data from exponential families, while reducing the dimensionality of the compressed representations. This important property was known to be satisfied by the Gaussian IB (Chechik et al., 2005) but not known for other distributions. The Gaussian case is self-dual in that sense. Generalizing this property to other exponential families was an open problem for many years; (iv) The exponential form of the optimal encoder-decoder pairs allows for the application to distributions with special symmetries, which can be naturally expressed in this form.

We conclude that the exponential Dual Information Bottleneck (dualExpIB) framework allows us now to approach problems previously considered intractable by the IB. Furthermore, it may allow better predictions to IB applications using finite samples or data with low internal dimensionality or special symmetries.

References

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. *ArXiv*, abs/1612.00410, 2016.

Lawrence D. Brown. Fundamentals of statistical exponential families with applications in statistical decision theory. *Lecture Notes-Monograph Series*, 9:i–279, 1986. ISSN 07492170. URL http://www.jstor.org/stable/4355554.

Gal Chechik, Amir Globerson, Naftali Tishby, and Yair Weiss. Information bottleneck for gaussian variables. *J. Mach. Learn. Res.*, 6:165–188, December 2005. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1046920.1046926.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, New York, NY, USA, 2006. ISBN 0471241954.

Domenico Felice and Nihat Ay. Divergence functions in information geometry. In Frank Nielsen and Frédéric Barbaresco (eds.), *Geometric Science of Information - 4th International Conference, GSI 2019, Toulouse, France, August 27-29, 2019, Proceedings*, volume 11712 of *Lecture Notes in Computer Science*, pp. 433–442. Springer, 2019. ISBN 978-3-030-26979-1. doi: 10.1007/978-3-030-26980-7_45. URL https://doi.org/10.1007/978-3-030-26980-7_45.

Ran Gilad-bachrach, Amir Navot, and Naftali Tishby. An information theoretic tradeoff between complexity and accuracy. In *In Proceedings of the COLT*, pp. 595–609. Springer, 2003.

Amir Globerson and Naftali Tishby. Sufficient dimensionality reduction. *Journal of Machine Learning Research*, 3(Mar):1307–1331, 2003.

E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, May 1957. doi: 10.1103/PhysRev.106.620. URL https://link.aps.org/doi/10.1103/PhysRev.106.620.

S. Kullback. *Information Theory and Statistics*. Wiley, New York, 1959.

Xiang Lisa Li and Jason Eisner. Specializing word embeddings (for parsing) by information bottleneck. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*, pp. 2744–2754, Hong Kong, November 2019. URL http://cs.jhu.edu/~jason/papers/#li-eisner-2019. Best Paper Award.

Shuang Ma, Daniel McDuff, and Yale Song. Unpaired image-to-speech synthesis with multimodal information bottleneck. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7598–7607, 2019.

Amichai Painsky and Gregory W. Wornell. Bregman Divergence Bounds and the Universality of the Logarithmic Loss. *arXiv e-prints*, art. arXiv:1810.07014, Oct 2018.

Sonali Parbhoo, Mario Wieser, and Volker Roth. Causal deep information bottleneck. *ArXiv*, abs/1807.02326, 2018.

Albert E. Parker, Tomáš Gedeon, and Alexander G. Dimitrov. Annealing and the rate distortion problem. In S. Becker, S. Thrun, and K. Obermayer (eds.), *Advances in Neural Information Processing Systems 15*, pp. 993–976. MIT Press, 2003. URL http://papers.nips.cc/paper/2264-annealing-and-the-rate-distortion-problem.pdf.

Elad Schneidman, Noam Slonim, Naftali Tishby, R deRuyter van Steveninck, and William Bialek. Analyzing neural codes using the information bottleneck method. *Advances in Neural Information Processing Systems, NIPS*, 2001.

Ohad Shamir, Sivan Sabato, and Naftali Tishby. Learning and generalization with the information bottleneck. *Theor. Comput. Sci.*, 411:2696–2711, 2010.

Noam Slonim, Nir Friedman, and Naftali Tishby. Multivariate information bottleneck. *Neural Computation*, 18(8):1739–1789, 2006. doi: 10.1162/neco.2006.18.8.1739. URL `https://doi.org/10.1162/neco.2006.18.8.1739`. PMID: 16771652.

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

J. R. Tredicce, G. L. Lippi, Paul Mandel, B. Charasse, A. Chevalier, and B. Picqu. Critical slowing down at a bifurcation. *American Journal of Physics*, 72(6):799–809, 2004. doi: 10.1119/1.1688783. URL `https://doi.org/10.1119/1.1688783`.

G. Tusnady and I. Csiszar. Information geometry and alternating minimization procedures. *Statistics & Decisions: Supplement Issues*, 1:205–237, 1984.

Noga Zaslavsky and Naftali Tishby. Deterministic annealing and the evolution of optimal information bottleneck representations. *Preprint*, 2019.

A Appendix

A.1 dualIB Mathematical formulation

As presented in §1.6 the dualIB is solved with respect to the full Markov chain (equation 6) in which we introduce the new variable, $\hat{\mathrm{Y}}$, the *predicted label*. Thus, in analogy to the IB we want to write the optimization problem in term of $\hat{\mathrm{Y}}$.

Developing the expected distortion we find:

$$\begin{aligned}
\mathbb{E}_{p_\beta(\mathrm{x},\hat{\mathrm{x}})}[d_{\mathrm{dualIB}}(\mathrm{x},\hat{\mathrm{x}})] &= \sum_{\mathrm{x},\hat{\mathrm{x}}} p_\beta(\mathrm{x},\hat{\mathrm{x}}) \sum_{\hat{\mathrm{y}}} p_\beta(\hat{\mathrm{y}} \mid \hat{\mathrm{x}}) \log \frac{p_\beta(\hat{\mathrm{y}} \mid \hat{\mathrm{x}})}{p(\hat{\mathrm{y}} \mid \mathrm{x})} \\
&= \sum_{\hat{\mathrm{x}},\hat{\mathrm{y}}} p_\beta(\hat{\mathrm{x}}) p_\beta(\hat{\mathrm{y}} \mid \hat{\mathrm{x}}) \log \frac{p_\beta(\hat{\mathrm{y}} \mid \hat{\mathrm{x}})}{p_\beta(\hat{\mathrm{y}})} - \sum_{\mathrm{x},\hat{\mathrm{y}}} p(\mathrm{x}) p_\beta(\hat{\mathrm{y}} \mid \mathrm{x}) \log \frac{p_\beta(\hat{\mathrm{y}} \mid \mathrm{x})}{p_\beta(\hat{\mathrm{y}})} \\
&+ \sum_{\mathrm{x},\hat{\mathrm{y}}} p(\mathrm{x}) p_\beta(\hat{\mathrm{y}} \mid \mathrm{x}) \log \frac{p_\beta(\hat{\mathrm{y}} \mid \mathrm{x})}{p(\hat{\mathrm{y}} \mid \mathrm{x})} \\
&= I(\hat{\mathrm{X}};\hat{\mathrm{Y}}) - I(\mathrm{X};\hat{\mathrm{Y}}) + \mathbb{E}_{p(\mathrm{x})}[D[p_\beta(\hat{\mathrm{y}} \mid \mathrm{x}) \| p(\hat{\mathrm{y}} \mid \mathrm{x})]].
\end{aligned}$$

Allowing the dual optimization problem to be written as:

$$\mathcal{F}^*[p(\hat{\mathrm{x}} \mid \mathrm{x}); p(\hat{\mathrm{x}}); p(\mathrm{y} \mid \hat{\mathrm{x}})] = I(\mathrm{X};\hat{\mathrm{X}}) - \beta\Big\{ I(\mathrm{X};\hat{\mathrm{Y}}) - I(\hat{\mathrm{X}};\hat{\mathrm{Y}}) - \mathbb{E}_{p(\mathrm{x})}[D[p_\beta(\hat{\mathrm{y}} \mid \mathrm{x}) \| p(\hat{\mathrm{y}} \mid \mathrm{x})]] \Big\}.$$

A.2 The DualIB solutions

To prove *theorem* 1 we want to obtain the normalized distributions minimizing the dualIB rate-distortion problem.

Proof. (i) Given that the problem is formulated as a rate-distortion problem the encoder's update rule must be the known minimizer of the distortion function (Cover & Thomas, 2006). Thus the IB encoder with the dual distortion is plugged in. (ii) For the decoder, by considering a small perturbation in the distortion $d_{\mathrm{dualIB}}(\mathrm{x},\hat{\mathrm{x}})$, with $\alpha(\hat{\mathrm{x}})$ the normalization Lagrange multiplier, we obtain:

$$\begin{aligned}
\delta d_{\mathrm{dualIB}}(\mathrm{x},\hat{\mathrm{x}}) &= \delta\left(\sum_{\mathrm{y}} p_\beta(\mathrm{y} \mid \hat{\mathrm{x}}) \log \frac{p_\beta(\mathrm{y} \mid \hat{\mathrm{x}})}{p(\mathrm{y} \mid \mathrm{x})} + \alpha(\hat{\mathrm{x}}) \left(\sum_{\mathrm{y}} p_\beta(\mathrm{y} \mid \hat{\mathrm{x}}) - 1 \right) \right) \\
\frac{\delta d_{\mathrm{dualIB}}(\mathrm{x},\hat{\mathrm{x}})}{\delta p_\beta(\mathrm{y} \mid \hat{\mathrm{x}})} &= \log \frac{p_\beta(\mathrm{y} \mid \hat{\mathrm{x}})}{p(\mathrm{y} \mid \mathrm{x})} + 1 + \alpha(\hat{\mathrm{x}}).
\end{aligned}$$

Hence, minimizing the expected distortion becomes:

$$\begin{aligned}
0 &= \sum_{\mathrm{x}} p_\beta(\mathrm{x} \mid \hat{\mathrm{x}}) \left[\log \frac{p_\beta(\mathrm{y} \mid \hat{\mathrm{x}})}{p(\mathrm{y} \mid \mathrm{x})} + 1 \right] + \alpha(\hat{\mathrm{x}}) \\
&= \log p_\beta(\mathrm{y} \mid \hat{\mathrm{x}}) - \sum_{\mathrm{x}} p_\beta(\mathrm{x} \mid \hat{\mathrm{x}}) \log p(\mathrm{y} \mid \mathrm{x}) + 1 + \alpha(\hat{\mathrm{x}}),
\end{aligned}$$

which yields Algorithm 1, row 6. □

Considering the dualIB encoder-decoder, Algorithm 1, we find that $\mathbb{E}_{p_\beta(\mathrm{x},\hat{\mathrm{x}})}[d_{\mathrm{dualIB}}(\mathrm{x},\hat{\mathrm{x}})]$ reduces to the expectation of the decoder's $\log$ partition function:

$$\begin{aligned}
\mathbb{E}_{p_\beta(\mathrm{x},\hat{\mathrm{x}})}[d_{\mathrm{dualIB}}(\mathrm{x},\hat{\mathrm{x}})] &= \sum_{\mathrm{x},\hat{\mathrm{x}}} p_\beta(\mathrm{x},\hat{\mathrm{x}}) \sum_{\mathrm{y}} p_\beta(\mathrm{y} \mid \hat{\mathrm{x}}) \log \frac{p_\beta(\mathrm{y} \mid \hat{\mathrm{x}})}{p(\mathrm{y} \mid \mathrm{x})} \\
&= -\mathbb{E}_{p_\beta(\hat{\mathrm{x}})}\left[\log Z_{\mathbf{y}|\hat{\mathbf{x}}}(\hat{\mathrm{x}};\beta)\right] + \sum_{\hat{\mathrm{x}},\mathrm{y}} p_\beta(\hat{\mathrm{x}}) \left[\sum_{\mathrm{x}'} p_\beta(\mathrm{x}' \mid \hat{\mathrm{x}}) \log p(\mathrm{y} \mid \mathrm{x}') - \sum_{\mathrm{x}} p_\beta(\mathrm{x} \mid \hat{\mathrm{x}}) \log p(\mathrm{y} \mid \mathrm{x}) \right] \\
&= -\mathbb{E}_{p_\beta(\hat{\mathrm{x}})}\left[\log Z_{\mathbf{y}|\hat{\mathbf{x}}}(\hat{\mathrm{x}};\beta)\right].
\end{aligned}$$

A.3 Stability analysis

Here we provide the detailed stability analysis allowing the definition of the matrices $C^{\mathrm{dualIB}}_{xx'}, C^{\mathrm{dualIB}}_{yy'}$ (equation 12) and which allows us to claim that they obey the same rules as the C matrices (equation 5) in equation 4. Considering a variation in $\hat{\mathrm{x}}$ we get:

$$
\begin{aligned}
\frac{\partial \log p_\beta(\mathrm{x} \mid \hat{\mathrm{x}})}{\partial \hat{\mathrm{x}}} &= \beta \sum_{\mathrm{y}} p_\beta(\mathrm{y} \mid \hat{\mathrm{x}}) \left(\log \frac{p(\mathrm{y} \mid \mathrm{x})}{p_\beta(\mathrm{y} \mid \hat{\mathrm{x}})} - 1 \right) \frac{\partial \log p_\beta(\mathrm{y} \mid \hat{\mathrm{x}})}{\partial \hat{\mathrm{x}}} \\
&= \beta \sum_{\mathrm{y}} p_\beta(\mathrm{y} \mid \hat{\mathrm{x}}) \left[\log p(\mathrm{y} \mid \mathrm{x}) - \sum_{\tilde{\mathrm{x}}} p_\beta(\tilde{\mathrm{x}} \mid \hat{\mathrm{x}}) \log p(\mathrm{y} \mid \tilde{\mathrm{x}}) \right] \frac{\partial \log p_\beta(\mathrm{y} \mid \hat{\mathrm{x}})}{\partial \hat{\mathrm{x}}} \\
&\quad + \beta \sum_{\mathrm{y}} \log Z_{\mathbf{y}|\hat{\mathbf{x}}}(\hat{\mathrm{x}}; \beta) \frac{\partial p_\beta(\mathrm{y} \mid \hat{\mathrm{x}})}{\partial \hat{\mathrm{x}}} \\
&= \beta \sum_{\mathrm{y}, \tilde{\mathrm{x}}} p_\beta(\mathrm{y} \mid \hat{\mathrm{x}}) p_\beta(\tilde{\mathrm{x}} \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y} \mid \mathrm{x})}{p(\mathrm{y} \mid \tilde{\mathrm{x}})} \frac{\partial \log p_\beta(\mathrm{y} \mid \hat{\mathrm{x}})}{\partial \hat{\mathrm{x}}}, \qquad (17)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \log p_\beta(\mathrm{y} \mid \hat{\mathrm{x}})}{\partial \hat{\mathrm{x}}} &= - \frac{1}{Z_{\mathbf{y}|\hat{\mathbf{x}}}(\hat{\mathrm{x}}; \beta)} \frac{\partial Z_{\mathbf{y}|\hat{\mathbf{x}}}(\hat{\mathrm{x}}; \beta)}{\partial \hat{\mathrm{x}}} + \sum_{\mathrm{x}} p_\beta(\mathrm{x} \mid \hat{\mathrm{x}}) \log p(\mathrm{y} \mid \mathrm{x}) \frac{\partial \log p_\beta(\mathrm{x} \mid \hat{\mathrm{x}})}{\partial \hat{\mathrm{x}}} \\
&= - \sum_{\tilde{\mathrm{y}}} p_\beta(\tilde{\mathrm{y}} \mid \hat{\mathrm{x}}) \sum_{\mathrm{x}} p_\beta(\mathrm{x} \mid \hat{\mathrm{x}}) \log p(\tilde{\mathrm{y}} \mid \mathrm{x}) \frac{\partial \log p_\beta(\mathrm{x} \mid \hat{\mathrm{x}})}{\partial \hat{\mathrm{x}}} \\
&\quad + \sum_{\mathrm{x}} p_\beta(\mathrm{x} \mid \hat{\mathrm{x}}) \log p(\mathrm{y} \mid \mathrm{x}) \frac{\partial \log p_\beta(\mathrm{x} \mid \hat{\mathrm{x}})}{\partial \hat{\mathrm{x}}} \\
&= \sum_{\mathrm{x}, \tilde{\mathrm{y}}} p_\beta(\mathrm{x} \mid \hat{\mathrm{x}}) p_\beta(\tilde{\mathrm{y}} \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y} \mid \mathrm{x})}{p(\tilde{\mathrm{y}} \mid \mathrm{x})} \frac{\partial \log p_\beta(\mathrm{x} \mid \hat{\mathrm{x}})}{\partial \hat{\mathrm{x}}}. \qquad (18)
\end{aligned}
$$

Substituting equation 18 into equation 17 and vice versa one obtains:

$$
\begin{aligned}
\frac{\partial \log p_\beta(\mathrm{x} \mid \hat{\mathrm{x}})}{\partial \hat{\mathrm{x}}} &= \beta \sum_{\mathrm{x}', \mathrm{y}, \tilde{\mathrm{y}}, \tilde{\mathrm{x}}} p_\beta(\mathrm{y} \mid \hat{\mathrm{x}}) p_\beta(\tilde{\mathrm{x}} \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y} \mid \mathrm{x})}{p(\mathrm{y} \mid \tilde{\mathrm{x}})} \\
&\qquad \cdot p_\beta(\mathrm{x}' \mid \hat{\mathrm{x}}) p_\beta(\tilde{\mathrm{y}} \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y} \mid \mathrm{x}')}{p(\tilde{\mathrm{y}} \mid \mathrm{x}')} \frac{\partial \log p_\beta(\mathrm{x}' \mid \hat{\mathrm{x}})}{\partial \hat{\mathrm{x}}} \\
\frac{\partial \log p_\beta(\mathrm{y} \mid \hat{\mathrm{x}})}{\partial \hat{\mathrm{x}}} &= \beta \sum_{\mathrm{x}, \mathrm{y}', \tilde{\mathrm{x}}, \tilde{\mathrm{y}}} p_\beta(\mathrm{x} \mid \hat{\mathrm{x}}) p_\beta(\tilde{\mathrm{y}} \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y} \mid \mathrm{x})}{p(\tilde{\mathrm{y}} \mid \mathrm{x})} \\
&\qquad \cdot p_\beta(\mathrm{y}' \mid \hat{\mathrm{x}}) p_\beta(\tilde{\mathrm{x}} \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y}' \mid \mathrm{x})}{p(\mathrm{y}' \mid \tilde{\mathrm{x}})} \frac{\partial \log p_\beta(\mathrm{y}' \mid \hat{\mathrm{x}})}{\partial \hat{\mathrm{x}}}.
\end{aligned}
$$

We now define the C^{dualIB} matrices as follows:

$$
\begin{aligned}
C^{\mathrm{dualIB}}_{\mathrm{xx}'}(\hat{\mathrm{x}}; \beta) &= \sum_{\mathrm{y}, \tilde{\mathrm{y}}, \tilde{\mathrm{x}}} p_\beta(\mathrm{y} \mid \hat{\mathrm{x}}) p_\beta(\tilde{\mathrm{x}} \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y} \mid \mathrm{x})}{p(\mathrm{y} \mid \tilde{\mathrm{x}})} \cdot p_\beta(\mathrm{x}' \mid \hat{\mathrm{x}}) p_\beta(\tilde{\mathrm{y}} \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y} \mid \mathrm{x}')}{p(\tilde{\mathrm{y}} \mid \mathrm{x}')} \\
C^{\mathrm{dualIB}}_{\mathrm{yy}'}(\hat{\mathrm{x}}; \beta) &= \sum_{\mathrm{x}, \tilde{\mathrm{x}}, \tilde{\mathrm{y}}} p_\beta(\mathrm{x} \mid \hat{\mathrm{x}}) p_\beta(\tilde{\mathrm{y}} \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y} \mid \mathrm{x})}{p(\tilde{\mathrm{y}} \mid \mathrm{x})} \cdot p_\beta(\mathrm{y}' \mid \hat{\mathrm{x}}) p_\beta(\tilde{\mathrm{x}} \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y}' \mid \mathrm{x})}{p(\mathrm{y}' \mid \tilde{\mathrm{x}})}.
\end{aligned}
$$

Using the above definition we have an equivalence to equation 4 in the form of:

$$
\left[I - \beta C^{\mathrm{dualIB}}_{\mathrm{xx}'}(\hat{\mathrm{x}}, \beta) \right] \frac{\partial \log p_\beta(\mathrm{x}' \mid \hat{\mathrm{x}})}{\partial \hat{\mathrm{x}}} = 0 \,, \quad \left[I - \beta C^{\mathrm{dualIB}}_{\mathrm{yy}'}(\hat{\mathrm{x}}, \beta) \right] \frac{\partial \log p_\beta(\mathrm{y}' \mid \hat{\mathrm{x}})}{\partial \hat{\mathrm{x}}} = 0.
$$

Note that for the binary case, the matrices may be simplified to:

$$
\begin{aligned}
C^{\mathrm{dualIB}}_{\mathrm{xx}'}(\hat{\mathrm{x}}; \beta) &= \sum_{\mathrm{y}, \tilde{\mathrm{x}}} p_\beta(\mathrm{y} \mid \hat{\mathrm{x}}) p_\beta(\tilde{\mathrm{x}} \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y} \mid \mathrm{x})}{p(\mathrm{y} \mid \tilde{\mathrm{x}})} \cdot p_\beta(\mathrm{x}' \mid \hat{\mathrm{x}}) (1 - p_\beta(\mathrm{y} \mid \hat{\mathrm{x}})) \log \frac{p(\mathrm{y} \mid \mathrm{x}')}{1 - p(\mathrm{y} \mid \mathrm{x}')} \\
C^{\mathrm{dualIB}}_{\mathrm{yy}'}(\hat{\mathrm{x}}; \beta) &= \sum_{\mathrm{x}, \tilde{\mathrm{x}}} p_\beta(\mathrm{x} \mid \hat{\mathrm{x}}) (1 - p_\beta(\mathrm{y} \mid \hat{\mathrm{x}})) \log \frac{p(\mathrm{y} \mid \mathrm{x})}{1 - p(\mathrm{y} \mid \mathrm{x})} \cdot p_\beta(\mathrm{y}' \mid \hat{\mathrm{x}}) p_\beta(\tilde{\mathrm{x}} \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y}' \mid \mathrm{x})}{p(\mathrm{y}' \mid \tilde{\mathrm{x}})}.
\end{aligned}
$$

A.3.1 Proof of Lemma 3

We show that the C^{dualIB} matrices share the same eigenvalues with $\lambda_1(\hat{\mathrm{x}}) = 0$.

Proof. The matrices, $C^{\text{dualIB}}_{\mathrm{xx}'}(\hat{\mathrm{x}};\beta)$, $C^{\text{dualIB}}_{\mathrm{yy}'}(\hat{\mathrm{x}};\beta)$, are given by:

$$C^{\text{dualIB}}_{\mathrm{xx}'}(\hat{\mathrm{x}};\beta) = A_{\mathrm{xy}}(\hat{\mathrm{x}};\beta)B_{\mathrm{yx}'}(\hat{\mathrm{x}};\beta)\ ,\ C^{\text{dualIB}}_{\mathrm{yy}'}(\hat{\mathrm{x}};\beta) = B_{\mathrm{yx}}(\hat{\mathrm{x}};\beta)A_{\mathrm{xy}'}(\hat{\mathrm{x}};\beta),$$

with:

$$A_{\mathrm{xy}}(\hat{\mathrm{x}};\beta) = p_\beta(\mathrm{y} \mid \hat{\mathrm{x}}) \sum_{\tilde{\mathrm{x}}} p_\beta(\tilde{\mathrm{x}} \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y} \mid \mathrm{x})}{p(\mathrm{y} \mid \tilde{\mathrm{x}})}\ ,\ B_{\mathrm{yx}}(\hat{\mathrm{x}};\beta) = p_\beta(\mathrm{x} \mid \hat{\mathrm{x}}) \sum_{\tilde{\mathrm{y}}} p_\beta(\tilde{\mathrm{y}} \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y} \mid \mathrm{x})}{p(\tilde{\mathrm{y}} \mid \mathrm{x})}.$$

Given that the matrices are obtained by the multiplication of the same matrices, it follows that they have the same eigenvalues $\{\lambda_i(\hat{\mathrm{x}};\beta)\}$.

To prove that $\lambda_1(\hat{\mathrm{x}};\beta) = 0$ we show that $\det(C^{\text{dualIB}}_{\mathrm{yy}'}) = 0$. We present the exact calculation for a binary label $\mathrm{Y} \in \{\mathrm{y}_0, \mathrm{y}_1\}$ (the argument for general Y follows by encoding the label as a sequence of bits and discussing the first bit only, as a binary case):

$$\begin{aligned}
\det(C^{\text{dualIB}}_{\mathrm{yy}'}(\hat{\mathrm{x}};\beta)) =& \sum_{\mathrm{x},\tilde{\mathrm{x}}} p_\beta(\mathrm{x} \mid \hat{\mathrm{x}}) p_\beta(\mathrm{y}_1 \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y}_0 \mid \mathrm{x})}{p(\mathrm{y}_1 \mid \mathrm{x})} \cdot p_\beta(\mathrm{y}_0 \mid \hat{\mathrm{x}}) p_\beta(\tilde{\mathrm{x}} \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y}_0 \mid \mathrm{x})}{p(\mathrm{y}_0 \mid \tilde{\mathrm{x}})} \\
&\cdot \sum_{\mathrm{x}',\tilde{\mathrm{x}}',} p_\beta(\mathrm{x}' \mid \hat{\mathrm{x}}) p_\beta(\mathrm{y}_0 \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y}_1 \mid \mathrm{x}')}{p(\mathrm{y}_0 \mid \mathrm{x}')} \cdot p_\beta(\mathrm{y}_1 \mid \hat{\mathrm{x}}) p_\beta\big(\tilde{\mathrm{x}}' \mid \hat{\mathrm{x}}\big) \log \frac{p(\mathrm{y}_1 \mid \mathrm{x}')}{p\big(\mathrm{y}_1 \mid \tilde{\mathrm{x}}'\big)} \\
&- \sum_{\mathrm{x},\tilde{\mathrm{x}}} p_\beta(\mathrm{x} \mid \hat{\mathrm{x}}) p_\beta(\mathrm{y}_0 \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y}_1 \mid \mathrm{x})}{p(\mathrm{y}_0 \mid \mathrm{x})} \cdot p_\beta(\mathrm{y}_0 \mid \hat{\mathrm{x}}) p_\beta(\tilde{\mathrm{x}} \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y}_0 \mid \mathrm{x})}{p(\mathrm{y}_0 \mid \tilde{\mathrm{x}})} \\
&\cdot \sum_{\mathrm{x}',\tilde{\mathrm{x}}'} p_\beta(\mathrm{x}' \mid \hat{\mathrm{x}}) p_\beta(\mathrm{y}_1 \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y}_0 \mid \mathrm{x}')}{p(\mathrm{y}_1 \mid \mathrm{x}')} \cdot p_\beta(\mathrm{y}_1 \mid \hat{\mathrm{x}}) p_\beta\big(\tilde{\mathrm{x}}' \mid \hat{\mathrm{x}}\big) \log \frac{p(\mathrm{y}_1 \mid \mathrm{x}')}{p\big(\mathrm{y}_1 \mid \tilde{\mathrm{x}}'\big)} \\
=& \sum_{\mathrm{x},\mathrm{x}',\tilde{\mathrm{x}},\tilde{\mathrm{x}}'} p_\beta(\mathrm{x} \mid \hat{\mathrm{x}}) p_\beta(\mathrm{x}' \mid \hat{\mathrm{x}}) p^2_\beta(\mathrm{y}_0 \mid \hat{\mathrm{x}}) p^2_\beta(\mathrm{y}_1 \mid \hat{\mathrm{x}}) p_\beta(\tilde{\mathrm{x}} \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y}_0 \mid \mathrm{x})}{p(\mathrm{y}_0 \mid \tilde{\mathrm{x}})} p_\beta\big(\tilde{\mathrm{x}}' \mid \hat{\mathrm{x}}\big) \log \frac{p(\mathrm{y}_1 \mid \mathrm{x}')}{p\big(\mathrm{y}_1 \mid \tilde{\mathrm{x}}'\big)} \\
&\cdot \left[\log \frac{p(\mathrm{y}_0 \mid \mathrm{x})}{p(\mathrm{y}_1 \mid \mathrm{x})} \log \frac{p(\mathrm{y}_1 \mid \mathrm{x}')}{p(\mathrm{y}_0 \mid \mathrm{x}')} - \log \frac{p(\mathrm{y}_0 \mid \mathrm{x})}{p(\mathrm{y}_1 \mid \mathrm{x})} \log \frac{p(\mathrm{y}_1 \mid \mathrm{x}')}{p(\mathrm{y}_0 \mid \mathrm{x}')}\right] = 0.
\end{aligned}$$

Given that the determinant is 0 implies that $\lambda_1(\hat{\mathrm{x}}) = 0$. □

For a binary problem we can describe the non-zero eigenvalue using $\lambda_2(\hat{\mathrm{x}}) = \mathrm{Tr}(C^{\text{dualIB}}_{\mathrm{yy}'}(\hat{\mathrm{x}};\beta))$. That is:

$$\begin{aligned}
\lambda_2(\hat{\mathrm{x}}) =& \sum_{\mathrm{x},\tilde{\mathrm{x}}} p_\beta(\mathrm{x} \mid \hat{\mathrm{x}}) p_\beta(\mathrm{y}_1 \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y}_0 \mid \mathrm{x})}{p(\mathrm{y}_1 \mid \mathrm{x})} \cdot p_\beta(\mathrm{y}_0 \mid \hat{\mathrm{x}}) p_\beta(\tilde{\mathrm{x}} \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y}_0 \mid \mathrm{x})}{p(\mathrm{y}_0 \mid \tilde{\mathrm{x}})} \\
&+ \sum_{\mathrm{x},\tilde{\mathrm{x}}} p_\beta(\mathrm{x} \mid \hat{\mathrm{x}}) p_\beta(\mathrm{y}_0 \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y}_1 \mid \mathrm{x})}{p(\mathrm{y}_0 \mid \mathrm{x})} \cdot p_\beta(\mathrm{y}_1 \mid \hat{\mathrm{x}}) p_\beta(\tilde{\mathrm{x}} \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y}_1 \mid \mathrm{x})}{p(\mathrm{y}_1 \mid \tilde{\mathrm{x}})} \\
=& p_\beta(\mathrm{y}_1 \mid \hat{\mathrm{x}}) p_\beta(\mathrm{y}_0 \mid \hat{\mathrm{x}}) \sum_{\mathrm{x},\tilde{\mathrm{x}}} p_\beta(\mathrm{x} \mid \hat{\mathrm{x}}) p_\beta(\tilde{\mathrm{x}} \mid \hat{\mathrm{x}}) \log \frac{p(\mathrm{y}_0 \mid \mathrm{x})}{p(\mathrm{y}_1 \mid \mathrm{x})} \left[\log \frac{p(\mathrm{y}_0 \mid \mathrm{x})}{p(\mathrm{y}_0 \mid \tilde{\mathrm{x}})} - \log \frac{p(\mathrm{y}_1 \mid \mathrm{x})}{p(\mathrm{y}_1 \mid \tilde{\mathrm{x}})}\right].
\end{aligned}$$

A.4 Information Plane analysis

We rely on known results for the rate-distortion problem and the information plane:

Lemma 8. *$I(\mathrm{X};\hat{\mathrm{X}})$ is a non-increasing convex function of the distortion $\mathbb{E}_{p_\beta(\mathrm{x},\hat{\mathrm{x}})}[d(\mathrm{x},\hat{\mathrm{x}})]$ with a slope of $-\beta$.*

We emphasis that this is a general result of rate-distortion thus holds for the dualIB as well.

Lemma 9. *For a fixed encoder $p_\beta(\hat{\mathrm{x}} \mid \mathrm{x})$ and the Bayes optimal decoder $p_\beta(\mathrm{y} \mid \hat{\mathrm{x}})$:*

$$\mathbb{E}[d_{\mathrm{IB}}(\mathrm{x}, \hat{\mathrm{x}})] = I(\mathrm{X}; \mathrm{Y}) - I(\hat{\mathrm{X}}; \mathrm{Y}).$$

Thus, the information curve, I_{Y} vs. I_{X}, is a non-decreasing concave function with a positive slope, β^{-1}. The concavity implies that β increases along the curve.

(Cover & Thomas, 2006; Gilad-bachrach et al., 2003).

A.4.1 Proof of Lemma 4

In the following section we provide a proof to *lemma* 4, for the IB and dualIB problems.

Proof. We want to analyze the behavior of $I_{\mathrm{X}}(\beta)$, $I_{\mathrm{Y}}(\beta)$, that is the change in each term as a function of the corresponding β. From *lemma* 9, the concavity of the information curve, we can deduce that both are non-decreasing functions of β. As the two β derivatives are proportional it's enough to discuss the first one.

Next, we focus on their behavior between two critical points. That is, where the cardinality of $\hat{\mathrm{X}}$ is fixed (clusters are ”static”). For ”static” clusters, the β derivative of I_{X}, along the optimal line is given by:

$$\begin{aligned}
\frac{\partial I(\mathrm{X}; \hat{\mathrm{X}})}{\partial \beta} &= -\frac{\partial}{\partial \beta}\left[\sum_{\mathrm{x}, \hat{\mathrm{x}}} p_\beta(\mathrm{x}, \hat{\mathrm{x}})\big(\log Z_{\hat{\mathbf{x}}|\mathbf{x}}(\mathrm{x}; \beta) + \beta d(\mathrm{x}, \hat{\mathrm{x}})\big)\right] \\
&= -\beta\left\langle d(\mathrm{x}, \hat{\mathrm{x}}) \frac{\partial \log p_\beta(\hat{\mathrm{x}} \mid \mathrm{x})}{\partial \beta}\right\rangle_{p_\beta(\mathrm{x}, \hat{\mathrm{x}})} \\
&\approx \beta\left\langle d(\mathrm{x}, \hat{\mathrm{x}})\left[\frac{\partial \log Z_{\hat{\mathbf{x}}|\mathbf{x}}(\mathrm{x}; \beta)}{\partial \beta} + d(\mathrm{x}, \hat{\mathrm{x}})\right]\right\rangle_{p_\beta(\mathrm{x}, \hat{\mathrm{x}})} \\
&\approx \beta\left\langle \underbrace{\left\langle d^2(\mathrm{x}, \hat{\mathrm{x}})\right\rangle_{p_\beta(\hat{\mathrm{x}}|\mathrm{x})} - \left\langle d(\mathrm{x}, \hat{\mathrm{x}})\right\rangle^2_{p_\beta(\hat{\mathrm{x}}|\mathrm{x})}}_{\mathrm{Var}(d(\mathrm{x}))}\right\rangle_{p(\mathrm{x})} .
\end{aligned}$$

This first of all reassures that the function is non-decreasing as $\mathrm{Var}(d(\mathrm{x})) \geq 0$.

The piece-wise concavity follows from the fact that when the number of clusters is fixed (between the critical points) - increasing β decreases the clusters conditional entropy $H(\hat{\mathrm{X}} \mid \mathrm{X})$, as the encoder becomes more deterministic. The mutual information is bounded by $H(\hat{\mathrm{X}})$ and it's β derivative decreases. Further, between the critical points there are no sign changes in the second β derivative. □

A.4.2 Proof of Lemma 5

Proof. The information curve has a positive slope, β^{-1}, with β increasing along it, *lemma* 9. That is, given a value of β, there exists a pair $I_{\mathrm{Y}}^{\mathrm{IB}}(\beta), I_{\mathrm{X}}^{\mathrm{IB}}(\beta)$ such that $\partial I_{\mathrm{Y}}^{\mathrm{IB}}(\beta)/\partial I_{\mathrm{X}}^{\mathrm{IB}}(\beta) = \beta^{-1}$. Now, consider a sub-optimal information curve, $I_{\mathrm{Y}}^*, I_{\mathrm{X}}^*$. There exist values β', β'' such that:

$$I_{\mathrm{X}}^*(\beta') = I_{\mathrm{X}}^{\mathrm{IB}}(\beta)\ ,\ \ I_{\mathrm{Y}}^*(\beta'') = I_{\mathrm{Y}}^{\mathrm{IB}}(\beta).$$

The optimality of the IB implies that sub-optimal curves lie below it; i.e, the IB slope is steeper:

$$\beta^{-1} > \beta'^{-1}\ ,\ \ \beta^{-1} > \beta''^{-1}.$$

Thus, given that β increases along the information curve, it holds that:

$$I_{\mathrm{X}}^*(\beta) > I_{\mathrm{X}}^*(\beta') = I_{\mathrm{X}}^{\mathrm{IB}}(\beta)\ ,\ \ I_{\mathrm{Y}}^*(\beta) > I_{\mathrm{Y}}^*(\beta'') = I_{\mathrm{Y}}^{\mathrm{IB}}(\beta).$$

□

A.5 Derivation of the dualExpIB

We provide elaborate derivations to *theorem* 7; that is, we obtain the dualIB optimal encoder-decoder under the exponential assumption over the data. We use the notations defined in §3.

- The *decoder*, equation 14.
  Substituting the exponential assumption into the dualIB log-decoder yields:

$$\begin{aligned}\log p_\beta(\mathrm{y}\mid\hat{\mathrm{x}}) &= \sum_{\mathrm{x}} p_\beta(\mathrm{x}\mid\hat{\mathrm{x}})\log p(\mathrm{y}\mid\mathrm{x}) - \log Z_{\mathbf{y}|\hat{\mathbf{x}}}(\hat{\mathrm{x}};\beta)\\ &= -\sum_{\mathrm{x}}\sum_{r=0}^{d} p_\beta(\mathrm{x}\mid\hat{\mathrm{x}})\lambda^r(\mathrm{y})A_r(\mathrm{x}) - \log Z_{\mathbf{y}|\hat{\mathbf{x}}}(\hat{\mathrm{x}};\beta)\\ &= -\sum_{r=1}^{d}\lambda^r(\mathrm{y})A_{r,\beta}(\hat{\mathrm{x}}) - \mathbb{E}_{p_\beta(\mathrm{x}|\hat{\mathrm{x}})}\left[\lambda^0_{\mathbf{x}}\right] - \log Z_{\mathbf{y}|\hat{\mathbf{x}}}(\hat{\mathrm{x}};\beta).\end{aligned}$$

  Taking a closer look at the normalization term:

$$\begin{aligned}Z_{\mathbf{y}|\hat{\mathbf{x}}}(\hat{\mathrm{x}};\beta) &= \sum_{\mathrm{y}} e^{\sum_{\mathrm{x}} p_\beta(\mathrm{x}|\hat{\mathrm{x}})\log p(\mathrm{y}|\mathrm{x})} = e^{-\mathbb{E}_{p_\beta(\mathrm{x}|\hat{\mathrm{x}})}\left[\lambda^0_{\mathbf{x}}\right]}\sum_{\mathrm{y}} e^{-\sum_{r=1}^{d}\lambda^r(\mathrm{y})A_{r,\beta}(\hat{\mathrm{x}})}\\ \log Z_{\mathbf{y}|\hat{\mathbf{x}}}(\hat{\mathrm{x}};\beta) &= -\mathbb{E}_{p_\beta(\mathrm{x}|\hat{\mathrm{x}})}\left[\lambda^0_{\mathbf{x}}\right] + \log\left(\sum_{\mathrm{y}} e^{-\sum_{r=1}^{d}\lambda^r(\mathrm{y})A_{r,\beta}(\hat{\mathrm{x}})}\right).\end{aligned}$$

  From which it follows that $\lambda^0_\beta(\hat{\mathrm{x}})$ is given by:

$$\lambda^0_\beta(\hat{\mathrm{x}}) = \log\left(\sum_{\mathrm{y}} e^{-\sum_{r=1}^{d}\lambda^r(\mathrm{y})A_{r,\beta}(\hat{\mathrm{x}})}\right),$$

  and we can conclude that the dualExpIB decoder takes the form:

$$\log p_\beta(\mathrm{y}\mid\hat{\mathrm{x}}) = -\sum_{r=0}^{d}\lambda^r(\mathrm{y})A_{r,\beta}(\hat{\mathrm{x}}).$$

- The *encoder*, equation 15.
  The core of the encoder is the dual distortion function which may now be written as:

$$\begin{aligned}d_{\mathrm{dualIB}}(\mathrm{x},\hat{\mathrm{x}}) &= \sum_{\mathrm{y}} p_\beta(\mathrm{y}\mid\hat{\mathrm{x}})\log\frac{p_\beta(\mathrm{y}\mid\hat{\mathrm{x}})}{p(\mathrm{y}\mid\mathrm{x})}\\ &= \sum_{\mathrm{y}} p_\beta(\mathrm{y}\mid\hat{\mathrm{x}})\left[\left(\lambda^0_{\mathbf{x}} - \lambda^0_\beta(\hat{\mathrm{x}})\right) + \sum_{r=1}^{d}\lambda^r(\mathrm{y})(A_r(\mathrm{x}) - A_{r,\beta}(\hat{\mathrm{x}}))\right]\\ &= \lambda^0_{\mathbf{x}} - \lambda^0_\beta(\hat{\mathrm{x}}) + \sum_{r=1}^{d}\lambda^r_\beta(\hat{\mathrm{x}})(A_r(\mathrm{x}) - A_{r,\beta}(\hat{\mathrm{x}})),\end{aligned}$$

  substituting this into the encoder's definition we obtain:

$$\begin{aligned}p_\beta(\hat{\mathrm{x}}\mid\mathrm{x}) &= \frac{p_\beta(\hat{\mathrm{x}})}{Z_{\hat{\mathbf{x}}|\mathbf{x}}(\mathrm{x};\beta)}e^{-\beta\left[\lambda^0_{\mathbf{x}} - \lambda^0_\beta(\hat{\mathrm{x}}) + \sum_{r=1}^{d}\lambda^r_\beta(\hat{\mathrm{x}})[A_r(\mathrm{x}) - A_{r,\beta}(\hat{\mathrm{x}})]\right]}\\ &= \frac{p_\beta(\hat{\mathrm{x}})e^{\beta\lambda^0_\beta(\hat{\mathrm{x}})}}{Z_{\hat{\mathbf{x}}|\mathbf{x}}(\mathrm{x};\beta)}e^{-\beta\sum_{r=1}^{d}\lambda^r_\beta(\hat{\mathrm{x}})[A_r(\mathrm{x}) - A_{r,\beta}(\hat{\mathrm{x}})]}.\end{aligned}$$