# Compression-Based Regularization with an Application to Multi-Task Learning

Matias Vera, *Student Member, IEEE,* Leonardo Rey Vega, *Member, IEEE,*
and Pablo Piantanida, *Senior Member, IEEE*

***Abstract*—This paper investigates, from information theoretic grounds, a learning problem based on the principle that any regularity in a given dataset can be exploited to extract compact features from data, i.e., using fewer bits than needed to fully describe the data itself, in order to build meaningful representations of a relevant content (multiple labels). We begin studying a *multi-task learning* (MTL) problem from the average (over the tasks) of miss-classification probability point of view and linking it with the popular *cross-entropy* criterion. Our approach allows an information theoretic formulation of a MTL problem as a supervised learning framework in which the prediction models for several related tasks are learned jointly from common representations to achieve better generalization performance. More precisely, our formulation of the MTL problem can be interpreted as an *information bottleneck* problem with side information at the decoder. Based on that, we present an iterative algorithm for computing the optimal trade-offs and and some of its convergence properties are studied. An important feature of this algorithm is to provide a natural safeguard against overfitting, because it minimizes the average risk taking into account a penalization induced by the model complexity. Remarkably, empirical results illustrate that there exists an optimal information rate minimizing the *excess risk* which depends on the nature and the amount of available training data. Applications to hierarchical text categorization and distributional word clusters are also investigated, extending previous works.**

***Index Terms*—Multi-task learning, Information bottleneck, Regularization, Arimoto-Blahut algorithm, Side information.**

## I. Introduction

The data deluge of the recent decades leads to new expectations for scientific discoveries from massive data in biology, particle physics, social media, safety and e-commerce. While mankind is drowning in data, a significant part of it is unstructured; hence it is difficult to discover relevant information. A common denominator in these novel scenarios is the challenge of representation learning: how to extract salient features or statistical relationships from data in order to build meaningful representations of the relevant content.

Statistical models are used to acquire knowledge from data by identifying relationships between variables that allows making predictions and assessing their accuracy. An essential feature of learning is the generalization capability, i.e., its ability to successfully apply rules extracted from previously seen data to characterize unseen data [1]. It is known that complex models tend to produce *overfitting*, i.e., represent the training data too accurately, therefore diminishing their ability to handle unseen data. To palliate this inconvenient, regularization methods include parameter penalization, noise injection, and averaging over multiple models trained with different sample sets. Nevertheless, it is not clear how to optimally control model complexity and therefore, this problem is an active research topic.

The information bottleneck (IB) method was introduced by Tishby [2] with the goal of extracting the relevant information that some signal provides about another one that is of interest. The IB can be formulated in terms of a noisy source coding problem [3] with a log-loss fidelity criterion. The information-theoretic characterization of the corresponding rate-distortion function follows from the noisy source coding problem. This quantity, which has a paramount importance in its own, has also found applications in the field of statistical learning [1], where the log-loss fidelity criterion is a common and popular cost function. The rate-distortion function is also related to a similarity measure in unsupervised learning/cluster analysis and has already demonstrated substantial performance improvement over standard supervised and unsupervised learning methods in a variety of important applications including compression, estimation, pattern recognition and classification, and statistical regression (see [4] and references therein).

This paper is concerned with an iterative algorithm for computing the rate-distortion function (or more precisely the rate-relevance function) for an IB problem with side information, and its relationship with multi-task learning (MTL).

### A. Related work

Witsenhausen and Wyner [5] were the first who studied an information-theoretic problem equivalent to IB and obtained an interesting characterization of its solution and several applications to source coding. Whereas the IB method, in the same way as it will be studied presented below, was introduced in [2] as a rate-distortion problem with a log-loss fidelity measure. Since then, it was applied to derive several clustering algorithms for a wide variety of applications such as: text classification [6], galaxy spectra classification [7], speaker recognition [8], among others. Further information-theoretic extensions of the IB were recently considered in [9]–[13].

The algorithm for the computation of the classical rate-distortion problem was developed independently by Arimoto [14] and Blahut [15] and it is widely known as the

The work of M. Vera was supported by a Peruilh Ph.D. grant from Facultad de Ingeniería - Universidad de Buenos Aires. The work of L. Rey Vega was supported by grant PIP11220150100578CO.

M. Vera is with the Facultad de Ingeniería - Universidad de Buenos Aires, Argentina (e-mail: mvera@fi.uba.ar).

L. Rey Vega is with CSC-CONICET and the Facultad de Ingeniería - Universidad de Buenos Aires, Argentina (e-mail: lrey@fi.uba.ar).

P. Piantanida is with CentraleSupélec - CNRS - Université Paris-Sud, France (e-mail: pablo.piantanida@centralesupelec.fr).

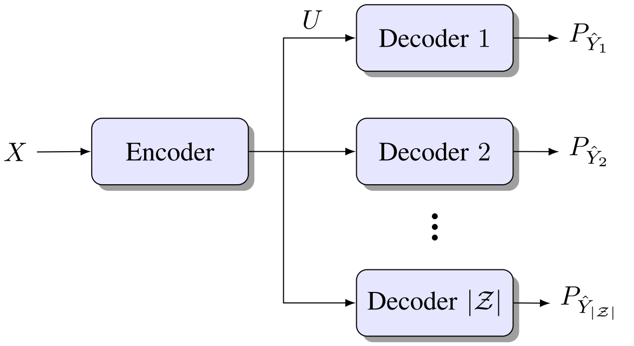

Fig. 1. MTL problem where $X$ denotes the input data, $U$ is the common feature (data representation) and the corresponding soft-estimates of the multiple labels are denoted by $P_{\hat{Y}_1|U}, \ldots, P_{\hat{Y}_{|\mathcal{Z}|}|U}$.

Blahut-Arimoto (BA) algorithm. An extension of this algorithm to the rate-distortion function with side information at the decoder was reported in [16]. Although this algorithm can be applied for optimizing the IB criterion [2], we emphasize that conventional algorithms [14], [15] are only expected to converge to a local minimum since IB leads to a non-convex problem due to the presence of the soft-encoder in the fidelity measure which depends on the optimizing distribution of the descriptions. Chechik *et al.* [17] adapts a BA algorithm to a restricted form of side information without further study of the involved optimization algorithm. In a different but related optimization problem, Kumar and Thangaraj [18] adapt the BA algorithm and analysis techniques provided in [19] to a non-convex problem while Yasui and Matsushima [20] extend this work for computing rate regions.

In this paper, we present a novel algorithm for MTL based on the IB paradigm with side information at the decoder. MTL is an approach to inductive transfer that improves generalization by using the domain information contained in the training signals of related tasks as an inductive bias [21]. This is accomplished by learning tasks in parallel while using a shared data representation, as described in Fig. 1. What is learned for each task can help other tasks to be learned better and thus can result in improved efficiency and prediction accuracy when compared to training the models separately [22]. MTL has received a great deal of attention in the recent years [23]. Application examples of MTL are e-mail classification (language recognition, topic recognition, spam or not), audio classification (speech recognition, speaker recognition, age verification), hierarchical text categorization (e.g. see Section IV-C), among others.

There are basically two ways of improving generalization via MTL. One approach imposes a structural condition on the learned parameters for all related tasks, e.g., by assuming some low-rank structure [24] or by modelling explicitly the links between tasks [25]. The other approach is through learning of common features for all desired tasks [26] via a common encoder (or feature selector) followed by a task-specific predictor, e.g., using a different decoder for each task. The later is the one we investigate in this paper. However, our setup differs from previous works in that we focus on an information-theoretic formulation of the MTL problem. We should also mention that we restrict our setup to MTL scenarios where the inputs are common to all tasks. Although this can be mathematically equivalent to the problem of multi-label learning (MLL), there are some important differences (see [27] for further details).

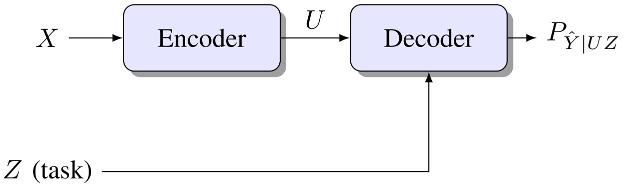

Fig. 2. MTL as the noisy source coding problem with log-loss fidelity and side information $Z$ (index task variable) at the decoder, referred to as the IB with side information.

### B. Our contribution

We introduce an information-theoretic paradigm which provides the fundamental trade-off between the *log-loss* (average risk) and the information rate of the features (statistical model complexity). We begin motivating our information theoretic approach to the MTL problem starting from the ultimate goal in a classification learning problem which is the misclassification probability and its relation with the cross-entropy risk. It worth to mention that our formulation, as the optimization of a rate-distortion function for a particular noisy source coding problem with side information at the decoder, provides an information-theoretic perspective for the MTL problem, which yields an intriguing connection between the fields of machine learning and information theory. Then, we derive an iterative Arimoto-Blahut like algorithm to address the non-convex optimization problem of the IB method in presence of side information available only at the decoder [9], [28], as described in Fig. 2, and consider several of its convergence properties. This approach provides a natural safeguard against overfitting by minimizing an average risk penalized by the model complexity. Remarkably, empirical results illustrate that there exists an optimal information rate minimizing the *excess risk* which depends on the nature and the amount of available training data. We further evaluate the performance of this algorithm on hierarchical text categorization of documents and numerical results demonstrates the merits of the proposed MTL algorithm in terms of the classification performance.

The rest of the paper is organized as follows. In Section II, we introduce the problem and present our iterative algorithm. The algorithm's properties are analyzed in Section III while in Section IV we show numerical evidence for some selected applications. Section V provides concluding remarks and major mathematical details are relegated to Appendices.

## II. Problem Definition and Main Result

### A. Notation and conventions

We use upper-case letters to denote random variables and lower-case letters to denote realizations of random variables

(RVs). Superscripts are used to denote the length of the vectors and subscripts denote the index of the components of a vector. All RVs live in finite alphabets. The probability mass function (pmf) of random variable $X$ is denoted by $P_X(x)$, $x \in \mathcal{X}$, where $\mathcal{X}$ is the alphabet of the random variable. When clear from the context we will simply refer to the pmf of $X$ as $P_X$. $\mathbb{E}_{P_X}[\cdot]$ denotes the expectation and $|\mathcal{A}|$ indicates the cardinality of a set $\mathcal{A}$. $A -\!\!\circ\!\!- B -\!\!\circ\!\!- C$ indicates a Markov chain, i.e., $P_{A|BC} = P_{A|B}$. The support of a pmf $P_X$ is denoted by $\text{supp}(P_X)$. The information measures to be used are [29]: the *entropy* $H(X) := \mathbb{E}_{P_X}[-\log P_X(X)]$, the *conditional entropy* $H(X|Y) := \mathbb{E}_{P_{X,Y}}\left[-\log P_{X|Y}(X|Y)\right]$ and the *relative entropy* or *Kullback Leibler divergence*:

$$\mathcal{D}(P_X \| Q_X) = \begin{cases} \mathbb{E}_{P_X}\left[\log \dfrac{P_X(X)}{Q_X(X)}\right] & \text{if } P_X \ll Q_X \\ +\infty & \text{otherwise,} \end{cases} \tag{1}$$

where we use $P_X \ll Q_X$ to denote that the probability measure $P_X$ is *absolutely continuous* w.r.t. $Q_X$, and the *mutual information*: $I(X;Y) := \mathcal{D}(P_{X,Y} \| P_X P_Y)$. When referring to an empirical distribution computed using data samples we will use notation $\hat{P}_X$. Functionals computed with an empirical distribution will be also denoted similarly, e.g., the entropy of $X$ computed by using $\hat{P}_X$ is denoted as: $\hat{H}(X)$. All logarithms are assumed to be base 2.

### B. Multi-Task Learning and Information Bottleneck

Consider a multi-task supervised classification problem. Let $(X, Y_1, \cdots, Y_{|\mathcal{Z}|})$ be arbitrary RVs. Soft-encoder $P_{U|X}$ is used to extract from $X$ information about a collection of labels $Y_z$ with $z \in \mathcal{Z} = \{1, \cdots, |\mathcal{Z}|\}$ (task index) and the corresponding soft-decoder $P_{\hat{Y}_z|U}$ seeks to recover the label for each task $z$, as shown in Fig. 1. In other words, the encoder aims to extracting relevant (common) information $U$ from a data set $X$ about the hidden labels $Y_z$ for each $z \in \mathcal{Z}$. This common information is used for solving the collection of classification tasks $z \in \mathcal{Z}$ at the decoder side. The misclassification probability for each task is defined as:

$$\mathbb{P}\big(Y_z \neq \hat{Y}_z\big) = 1 - \mathbb{E}_{P_{X,Y_z} P_{U|X}}\left[P_{\hat{Y}_z|U}(Y_z|U)\right]. \tag{2}$$

The MTL goal is to achieve low misclassification probability for all tasks. But is clear that depending on the application some task might be more relevant than others. It is reasonable to define an artificial random variable $Z$, whose pmf $P_Z(z)$ represents the relative importance of each task $z \in \mathcal{Z}$. The problem, is then shown in Fig. 2, where the variable $Y$ is defined such that $P_{X,Y|Z}(x,y|z) = P_{X,Y_z}(x,y)$ and the soft-decoder can be written as $P_{\hat{Y}|U,Z}(y|u,z) = P_{\hat{Y}_z|U}(y|u)$. The encoder is given as before in terms of $P_{U|X}$, given the fact that we enforce the description $U$ to be common to all tasks. In this way, the average misclassification probability is given by $\mathbb{P}\big(Y \neq \hat{Y}\big) = \mathbb{E}_{P_Z}[\mathbb{P}\big(Y_z \neq \hat{Y}_z\big)]$. This probability has the particularity that it is mathematically hard to optimize. As a consequence, it is common to work with a surrogate function given by the *logarithmic loss* $\log P_{\hat{Y}|U,Z}(Y|U,Z)$ which give rise to the average *cross-entropy* risk:

$$\text{Risk}\big(P_{U|X}, P_{\hat{Y}|U,Z}\big) := \mathbb{E}_{P_{X,Y,Z} P_{U|X}}[-\log P_{\hat{Y}|U,Z}(Y|U,Z)] \tag{3}$$

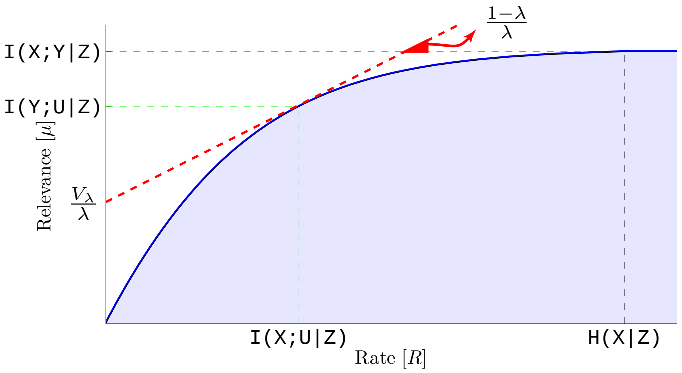

Fig. 3. A relevance-rate region with its supporting hyperplane (9).

It is straightforward to obtain that:

$$\mathbb{P}\big(Y \neq \hat{Y}\big) \leq 1 - 2^{-\text{Risk}\left(P_{U|X}, P_{\hat{Y}|U,Z}\right)}, \tag{4}$$

which shows that the cross-entropy provides an upper bound to the average misclassification probability, which motivates its used as surrogate cost function for the MTL problem. Besides this observation, cross-entropy risk provides in general an effective and better behaved cost for minimizing the misclassification probability. An interesting observation is given by the fact that in the above problem the optimal choice for the decoder $P_{\hat{Y}|U,Z}(Y|U,Z)$ becomes:

$$P_{Y|U,Z}(y|u,z) = \frac{\sum_x P_{U|X}(u|x) P_{X,Y,Z}(x,y,z)}{\sum_x P_{U|X}(u|x) P_{X,Z}(x,z)}, \tag{5}$$

which is completely determined by the encoder $P_{U|X}$ and input distribution. With this choice of the decoder, the risk is given by $\text{Risk}\big(P_{U|X}\big) = H(Y|U,Z)$. As a consequence the problem is reduced to the one of finding the encoder $P_{U|X}$ which minimizes $H(Y|U,Z)$ or equivalent, of maximizing mutual information $I(Y;U|Z)$.

The representation $U$ is expected to summarize data $X$ in a compact way, where compactness of the model is measured in terms of the minimum Shannon entropy rate. However, learning a representation $U$ for predicting $Y$ requires to capture the regularities in $Y$ that are present in $X$ while other irrelevant information for $Y$ must be disregarded. In this sense, our statistical measure of complexity says that the best description $U$ of the data is given by a model able to compress $X$ which is captured by Shannon mutual-information rate $I(U;X|Z=z)$. In the spirit of the Kolmogorov-Chaitin complexity [30] this is a measure of the regularities present in an object above and beyond pure randomness. To summarize finding the encoder that minimizes the average log-loss or cross-entropy is equivalent to search for the encoder maximizing the mutual (relevance) information $I(Y;U|Z) = H(Y|Z) - H(Y|U,Z)$. As a consequence, we can focus on maximizing the relevance (mutual information) $I(Y;U|Z)$ subject to a given complexity (Shannon rate) $I(X;U|Z)$.

*Definition 1 (Relevance-rate region):* A pair rates $(R,\mu)$ is achievable iff it belongs to the rate-relevance region:

$$\begin{aligned} \mathcal{R} := \Big\{(\mu, R) \in \mathbb{R}^2_{\geq 0} : \exists\, P_{U|X} \text{ s.t. } \; & R \geq I(U;X|Z), \\ & \mu \leq I(Y;U|Z), \quad U -\!\!\circ\!\!- X -\!\!\circ\!\!- (Y,Z)\Big\}. \end{aligned} \tag{6}$$

and the corresponding relevance-rate function is defined by

$$L(R, P_{X,Y,Z}) = \max \{\mu : (R, \mu) \in \mathcal{R}\}, \quad (7)$$

$$= \max_{P_{U|X}: I(X;U|Z) \le R} I(Y;U|Z). \quad (8)$$

The computation of this function can be interpreted as an IB problem with side information only at the decoder. At this point, we should mention that problem (8) was obtained from a single letter formulation of the MTL problem in terms of an encoder which generates common representations and a family of decoders which perform the classification tasks. Although, it has been shown in [9, Theo. 1 with $\mu_2 = 0$, $L = 1$] that the above problem has also an operational significance in information-theoretic terms quantifying the asymptotic (with increasing block-lengths) trade-offs between a multi-letter relevance and the compression rate, this interpretation, however has not major significance for the learning problem where inputs are generally treated in single-letter basis and there is no block-length tending to infinity.

It is worth to mention that $L_C(R, P_{X,Y}) \le L(R, P_{X,Y,Z}) \le L_{SED}(R, P_{X,Y,Z})$, where $L_C(R, P_{X,Y})$ refers to the classical IB relevance-rate function and $L_{SED}(R, P_{X,Y,Z})$ to the one with side information at both the encoder and the decoder. This behaviour is related to the fact that any additional knowledge can generate relevant information [11]. Interestedly, these two problems are particular cases of (8): when $|\mathcal{Z}| = 1$ since the IB problem with side information becomes the standard one, while if $\tilde{X} = (X, Z)$, $L(R, P_{\tilde{X},Y,Z})$ becomes an IB problem with side information at both the encoder and the decoder. In this way, the problem in (8) is the most general and interesting. The following lemma is important for the characterization of the relevance-rate region:

*Lemma 1:* $\mathcal{R}$ is closed, convex and the cardinality of random variable $U$ can be bounded as $|\mathcal{U}| \le |\mathcal{X}| + 1$ without loss of generality.

The proof of this result is not difficult and for that reason is omitted. Observe that the relevance-rate function –as being the upper-boundary of $\mathcal{R}$– provides an alternative and complete characterization of the region. It is important to mention that the maximum in that problem is well-defined because we are attempting to maximize a continuous function over a compact set. A graphical example of the relevance-rate region can be seen in Fig. 3. Although the optimization involved in (8) does not lead to a convex problem, the properties of $\mathcal{R}$ allows us to characterize the optimal trade-off between compression and relevance rates using *supporting hyperplanes* [31]. As it is well known, any closed and convex set can be characterized by all its supporting hyperplanes [32]. A supporting hyperplane for $\mathcal{R}$ with parameter $\lambda$ can be written as:

$$V_\lambda := \max_{P_{U|X}} \lambda I(Y;U|Z) - (1-\lambda) I(X;U|Z). \quad (9)$$

With little effort it is easy to show that $\lambda \in [0,1]$ suffices for the full characterization of $\mathcal{R}$ using supporting hyperplanes. Finding the optimal encoder $P^{*,\lambda}_{U|X}$ in (9) requires knowledge of the underlying distribution $P_{X,Y,Z}$. In practical applications, this lack of knowledge is overcome by resorting to labeled examples, i.e., a training set of $n$ i.i.d. tuples: $\{(x_1, y_1, z_1), \dots, (x_n, y_n, z_n)\}$ sampled according to the unknown distribution $P_{X,Y,Z}$. In Section IV, we will study some supervised learning setups where expression (9) together with the iterative algorithm described below using empirical distribution $\hat{P}_{X,Y,Z}$ will serve as a supervised objective to guide MTL. Obviously, there other alternative methods which do not require the plug-in estimator but use an estimate of the source distribution, e.g., [33]–[35].

### C. An iterative optimization algorithm

In order to simplify the notation, we define $f(\lambda, P_{U|X})$ as:

$$f(\lambda, P_{U|X}) := \lambda I(Y;U|Z) - (1-\lambda) I(X;U|Z). \quad (10)$$

Clearly, we can write

$$f(\lambda, P_{U|X}) = \sum_{z \in \mathcal{Z}} P_Z(z) \big[\lambda I(Y;U|Z=z) - (1-\lambda) I(X;U|Z=z)\big], \quad (11)$$

where we see the effect of the weights $P_Z(z)$ associated with each task. *Data Processing Inequality* [29, sec. 2.3] allows to conclude that the only allowable values for the relevance are $0 \le \mu \le I(X;Y|Z)$. We wish to obtain an algorithm that is able to find the supporting hyperplanes of $\mathcal{R}$, for every $\lambda \in [0,1]$, allowing the computation of the upper-boundary of $\mathcal{R}$, i.e., finding the optimal pmf $P^{*,\lambda}_{U|X}$ that achieves the maximum in (9)

$$P^{*,\lambda}_{U|X} := \arg\max_{P_{U|X}} f(\lambda, P_{U|X}). \quad (12)$$

Using the Markov chain $U -\!\!\circ\!\!- X -\!\!\circ\!\!- (Y,Z)$, the function $f(\lambda, P_{U|X})$ can be written as:

$$f(\lambda, P_{U|X}) = (2\lambda - 1) I(X;U|Z) - \lambda I(X;U|Y,Z). \quad (13)$$

Depending on the value of $\lambda$, it is appropriate to define the algorithm in two different ways. This is similar to the approach in [20]. If $\lambda \in [0, 0.5]$, both terms of (13) are non-positive and thus, the solution is trivial: $V_\lambda = 0$. This is achieved for all pmf that satisfies $P_{U|X} = P_U$ and corresponds to the point $(0,0)$ in $\mathcal{R}$. The interesting case is when $\lambda \in (0.5, 1]$. In this case, the proposed iterative algorithm is summarized in Alg. 1, where $k_x$ are constants such that $\sum_u P^{(n+1)}_{U|X} = 1 \ \forall x \in \mathcal{X}$. In the next section, we explain the rationale behind this algorithm.

## III. Algorithm Analysis

The above problem is not convex because $P_{U|X} \mapsto f(\lambda, P_{U|X})$ is not concave. As a consequence, we cannot expect to devise an efficient procedure determining the global maximum of the problem. The algorithm proposed is a variant of the BA algorithm [14], [15] which is based on solid theoretical grounds and guarantee global optimum convergence results when the optimization problem is convex. Although the involved optimization problem is not convex, we provide some results regarding the convergence properties of the proposed algorithm. Convergence to a local maximum is guaranteed which could be also the global maximum provided that some theoretical conditions are fulfilled. These results are inspired from seminal works in [18]–[20].

**ALGORITHM 1:** Information Bottleneck with side information.

```
Input: $P_{X,Y,Z}$, $P^{(0)}_{U|X}$, $\lambda \in [0,1]$, $\epsilon > 0$.
Output: $P^{*,\lambda}_{U|X}$.
  Initialize $n := 0$, $I^{(0)}_{\lambda} := +\infty$, $F^{(0)}_{\lambda} := 0$.
  while $I^{(n)}_{\lambda} - F^{(n)}_{\lambda} > \epsilon$ do
    Compute
      $Q^{(n+1)}_{U|Y,Z} := \sum_{x} P^{(n)}_{U|X} P_{X|Y,Z}, \quad Q^{(n+1)}_{X|Z,U} := \frac{P^{(n)}_{U|X} P_{X|Z}}{\sum_{x'} P^{(n)}_{U|X'} P_{X'|Z}},$
      $P^{(n+1)}_{U|X} := k_x \cdot \exp\left\{ \frac{2\lambda-1}{\lambda} \sum_{z} P_{Z|X} \log(Q^{(n+1)}_{X|Z,U}) + \sum_{y,z} P_{Y|X,Z} P_{Z|X} \log(Q^{(n+1)}_{U|Y,Z}) \right\},$
      $F^{(n+1)}_{\lambda} := (2\lambda-1) \sum_{x,z,u} P^{(n+1)}_{U|X} P_{X,Z} \log\left(\frac{Q^{(n+1)}_{X|Z,U}}{P_{X|Z}}\right) - \lambda \sum_{x,y,z,u} P^{(n+1)}_{U|X} P_{X,Y,Z} \log\left(\frac{P^{(n+1)}_{U|X}}{Q^{(n+1)}_{U|Y,Z}}\right),$
      $I^{(n+1)}_{\lambda} := \max_{u} (2\lambda-1) \sum_{x,z} P_{X,Z} \log\left(\frac{Q^{(n+1)}_{X|Z,U}}{P_{X|Z}}\right) - \lambda \sum_{x,y,z} P_{X,Y,Z} \log\left(\frac{P^{(n)}_{U|X}}{Q^{(n+1)}_{U|Y,Z}}\right),$
    Update $n := n+1$.
  end while
  Report $P^{*,\lambda}_{U|X} = P^{(n)}_{U|X}$.
```

### *A. Algorithm summary*

We first study the algorithms expressions in further detail. Eq. (13) can be expanded as:

$$f(\lambda, P_{U|X}) = (2\lambda-1) \sum_{x,z,u} P_{U|X} P_{X,Z} \log\left(\frac{P_{X|Z,U}}{P_{X|Z}}\right) - \lambda \sum_{x,y,z,u} P_{U|X} P_{X,Y,Z} \log\left(\frac{P_{U|X}}{P_{U|Y,Z}}\right). \quad (14)$$

Let the function $F(\lambda, P_{U|X}, Q_{U|Y,Z}, Q_{X|Z,U})$ be:

$$\begin{aligned} &F(\lambda, P_{U|X}, Q_{U|Y,Z}, Q_{X|Z,U}) \\ &\quad := (2\lambda-1) \sum_{x,z,u} P_{U|X} P_{X,Z} \log\left(\frac{Q_{X|Z,U}}{P_{X|Z}}\right) \\ &\quad - \lambda \sum_{x,y,z,u} P_{U|X} P_{X,Y,Z} \log\left(\frac{P_{U|X}}{Q_{U|Y,Z}}\right), \end{aligned} \quad (15)$$

where $Q_{U|Y,Z}, Q_{X|Z,U}$ are arbitrary pmfs. For sake of simplicity, sometimes we write $F$ when the arguments are obvious. This new function has some important properties.

*Lemma 2:* Consider any $P_{U|X}$ and let $\lambda \in (0.5, 1]$. The following properties hold true:

1) $f(\lambda, P_{U|X}) \geq F(\lambda, P_{U|X}, Q_{U|Y,Z}, Q_{X|Z,U})$, and equality is achieved iff $Q_{U|Y,Z} = P_{U|Y,Z} \ \forall\ (y,z) \in \mathcal{Y} \times \mathcal{Z}$ and $Q_{X|Z,U} = P_{X|Z,U}\ \forall (z,u) \in \mathcal{Z} \times \mathcal{U}$.
2) The value $V_\lambda$ satisfies:

$$V_\lambda = \max_{P_{U|X}} \max_{Q_{U|Y,Z}, Q_{Y|Z,U}} F(\lambda, P_{U|X}, Q_{U|Y,Z}, Q_{X|Z,U}). \quad (16)$$

3) For any $Q_{U|Y,Z}, Q_{X|Z,U}$ and $\lambda \in (0.5, 1]$, $P_{U|X} \mapsto F(\lambda, P_{U|X}, Q_{U|Y,Z}, Q_{X|Z,U})$ is concave and achieves its maximum provided that:

$$\begin{aligned} P_{U|X} = k_x \cdot \exp\Big\{ &\frac{2\lambda-1}{\lambda} \sum_{z \in \mathcal{Z}} P_{Z|X} \log(Q_{X|Z,U}) \\ &+ \sum_{(y,z) \in \mathcal{Y} \times \mathcal{Z}} P_{Y|X,Z} P_{Z|X} \log(Q_{U|Y,Z}) \Big\}, \end{aligned} \quad (17)$$

where $k_x$ are constants such that $\sum_u P_{U|X} = 1\ \forall x \in \mathcal{X}$.

*Proof:*

1) The difference between functions can be written as:

$$\begin{aligned} &f(\lambda, P_{U|X}) - F(\lambda, P_{U|X}, Q_{U|Y,Z}, Q_{X|Z,U}) \\ &\quad = (2\lambda-1) \sum_{x,y,z,u} P_{U|X} P_{X,Y,Z} \log\left(\frac{P_{X|Z,U}}{P_{X|Z}}\right) \\ &\quad\quad - \lambda \sum_{x,y,z,u} P_{U|X} P_{X,Y,Z} \log\left(\frac{P_{U|X}}{P_{U|Y,Z}}\right) \\ &\quad\quad - (2\lambda-1) \sum_{x,y,z,u} P_{U|X} P_{X,Y,Z} \log\left(\frac{Q_{X|Z,U}}{P_{X|Z}}\right) \\ &\quad\quad + \lambda \sum_{x,y,z,u} P_{U|X} P_{X,Y,Z} \log\left(\frac{P_{U|X}}{Q_{U|Y,Z}}\right) \quad (18) \\ &\quad = \lambda \sum_{y,z} P_{Y,Z}\ \mathcal{D}(P_{U|Y,Z} \| Q_{U|Y,Z}) \\ &\quad\quad + (2\lambda-1) \sum_{z,u} P_{Z,U}\ \mathcal{D}(P_{X|Z,U} \| Q_{X|Z,U}) \geq 0 \end{aligned} \quad (19)$$

with equality iff $Q_{U|Y,Z} = P_{U|Y,Z}\ \forall\ (y,z) \in \mathcal{Y} \times \mathcal{Z}$ and $Q_{X|Z,U} = P_{X|Z,U}\ \forall\ (z,u) \in \mathcal{Z} \times \mathcal{U}$. This is easily seen from the properties of relative entropy [29, sec. 2.3].

2) The claim follows by combining the previous claim with (9).
3) Every pmf satisfies $\sum_u P_{U|X} = 1$. Then, using Lagrange multipliers $c_x,\ x \in \mathcal{X}$:

$$\frac{\partial[F + \sum_x c_x(\sum_u P_{U|X} - 1)]}{\partial P_{U|X}} = c_x - \lambda P_X \log(eP_{U|X}) + (2\lambda - 1)\sum_z P_{X,Z} \log\left(Q_{X|Z,U}\right) + \lambda \sum_{y,z} P_{X,Y,Z} \log(Q_{U|Y,Z}) = 0 \tag{20}$$

from which we immediately recover (17). Note that this solution meet $P_{U|X=x}(u) \geq 0$ for all $(x,u) \in \mathcal{X} \times \mathcal{U}$. The concavity results follow from:

$$\frac{\partial^2 F(\lambda, P_{U|X}, Q_{U|Y,Z}, Q_{Y|Z,U})}{\partial P^2_{U|X}} = -\frac{\lambda P_X \log(e)}{P_{U|X}} \leq 0. \tag{21}$$

■

We observe that the function $F(\lambda, P_{U|X}, Q_{U|Y,Z}, Q_{X|Z,U})$ provides an achievable and easy way to optimize a lower bound to the objective function $f(\lambda, P_{U|X})$, for each $P_{U|X}$. Interestingly, $P_{U|X} \mapsto F(\lambda, P_{U|X}, Q_{U|Y,Z}, Q_{X|Z,U})$ is concave for each $(Q_{U|Y,Z}, Q_{X|Z,U})$, guaranteeing that any local optimum is also a global one. These facts lead naturally to the iterative process in order to perform the double maximization which results in $V_\lambda$. This is the case in Algorithm 1, where we perform an iterative maximization process on both arguments: $P_{U|X}$ and $(Q_{U|Y,Z}, Q_{X|Z,U})$. For a given $\lambda \in (0.5, 1]$, starting from an initial condition $P^{(0)}_{U|X}$, and according to 2) in Lemma 2, we search for $Q^{(1)}_{U|Y,Z}, Q^{(1)}_{X|Z,U}$ such that the maximum of $F(\lambda, P^{(0)}_{U|X}, Q_{U|Y,Z}, Q_{X|Z,U})$ is achieved, for fixed $P^{(0)}_{U|X}$. Next, from 3) in the previous lemma, we find $P^{(1)}_{U|X}$ as the argument that maximizes $F(\lambda, P_{U|X}, Q^{(1)}_{U|Y,Z}, Q^{(1)}_{X|Z,U})$. This iterative process is repeated until a stopping criterion is satisfied (see Section III-C). It is easy to show that the sequence of values $F(\lambda, P^{(n)}_{U|X}, Q^{(n)}_{U|Y,Z}, Q^{(n)}_{X|Z,U})$ is monotone non-decreasing. This clearly guarantees the convergence. In the sequel, we further study this process in detail.

### *B. Convergence properties*

For sake of simplicity, let us assume that the optimal point $P^{*,\lambda}_{U|X}$ is unique. Define $F^{(n)}_\lambda := F(\lambda, P^{(n)}_{U|X}, Q^{(n)}_{U|Y,Z}, Q^{(n)}_{X|Z,U})$. From the previous section we know that $F^{(n+1)}_\lambda \geq F^{(n)}_\lambda$. Moreover, from 2) in Lemma 2, $V_\lambda \geq F^{(n)}_\lambda$ for all $n$. However, there is no guarantee that $V_\lambda = F^{(\infty)}_\lambda$. In order to obtain some insights on the convergence process and on the limiting point of the iterative process, we will consider the concept of $\delta$-superlevel set (see [18] for further details).

*Definition 2:* The $\delta$-superlevel set is defined as the set:

$$G_{\delta,\lambda} := \left\{ P_{U|X} : \mathcal{X} \to \mathcal{P}(\mathcal{U}) \middle|\ f(\lambda, P_{U|X}) \geq \delta \right\}. \tag{22}$$

*Definition 3:* Consider a fixed conditional distribution $\tilde{P} \in G_{\delta,\lambda}$. The set $H_{\delta,\lambda}(\tilde{P})$ is defined as the set of all points $P_{U|X} \in G_{\delta,\lambda}$ such that each of them (and $\tilde{P}$) are in the same path-connected component of $G_{\delta,\lambda}$. In order words, $H_{\delta,\lambda}(\tilde{P})$ is the set of all points $P_{U|X} \in G_{\delta,\lambda}$ that are reachable from $\tilde{P}$ by a continuous path.

*Lemma 3:* Let $\lambda \in (0.5, 1]$, the distribution $P^{(n+1)}_{U|X}$ lies in $H_{\delta,\lambda}(P^{(n)}_{U|X})$ for all $k$ such that $P^{(n)}_{U|X} \in G_{\delta,\lambda}$.

*Proof:* Let $\tilde{G}^n_{\delta,\lambda}$ be the $k$-superlevel of the function $F(\lambda, P_{U|X}, Q^{(n+1)}_{U|Y,Z}, Q^{(n+1)}_{X|Z,U})$. Since $f(\lambda, P_{U|X}) \geq F(\lambda, P_{U|X}, Q^{(n+1)}_{U|Y,Z}, Q^{(n+1)}_{X|Z,U})$ from Lemma 2 (i.e. by claim 1), it follows that $\tilde{G}^n_{\delta,\lambda} \subseteq G_{\delta,\lambda}\ \forall n$. Also, $P^{(n)}_{U|X}$ and $P^{(n+1)}_{U|X}$ lies in $\tilde{G}^n_{\delta,\lambda}$ because:

$$F^{(n+1)}_\lambda \geq F(\lambda, P^{(n)}_{U|X}, Q^{(n+1)}_{U|Y,Z}, Q^{(n+1)}_{X|Z,U}) = f(\lambda, P^{(n)}_{U|X}). \tag{23}$$

For fixed $(Q^{(n+1)}_{U|Y,Z}, Q^{(n+1)}_{X|Z,U})$ pmfs, we know that $F$ is concave in argument $P_{U|X}$. Thus, $\tilde{G}^n_{\delta,\lambda}$ is a convex set and it is therefore path-connected and between any two of its points there exists a continuous path. Then, it follows that $\tilde{G}^n_{\delta,\lambda} \subseteq H_{\delta,\lambda}(P^{(n)}_{U|X})$ and we conclude that $P^{(n+1)}_{U|X} \in H_{\delta,\lambda}(P^{(n)}_{U|X})$. ■

Clearly, this lemma and Definition 3 imply that if $P^{(0)}_{U|X} \in G_{\delta,\lambda}$ for a given value of $\delta$, then $P^{(n)}_{U|X} \in H_{\delta,\lambda}(P^{(0)}_{U|X})\ \forall n$ and the complete trajectory of the algorithm for a particular initial condition is contained in $H_{\delta,\lambda}(P^{(0)}_{U|X})$ which is clearly a path-connected set.

*Lemma 4:* Consider $\lambda \in (0.5, 1]$ and $P^{(0)}_{U|X} \in G_{\delta,\lambda}$ for a given value[1] of $\delta$. If the optimal solution $P^{*,\lambda}_{U|X}$ lies in $H_{\delta,\lambda}(P^{(0)}_{U|X})$, the function $f(\lambda, P_{U|X})$ is concave in $H_{\delta,\lambda}(P^{(0)}_{U|X})$, then the following inequalities hold for every $n$:

$$V_\lambda \leq \lambda \sum_{x,y,z,u} P^{*,\lambda}_{U|X} P_{X,Y,Z} \log\left(\frac{Q^{(n+1)}_{U|Y,Z}}{P^{(n)}_{U|X}}\right) + (2\lambda - 1) \sum_{x,z,u} P^{*,\lambda}_{U|X} P_{X,Z} \log\left(\frac{Q^{(n+1)}_{X|Z,U}}{P_{X|Z}}\right), \tag{24}$$

$$V_\lambda - F^{(n+1)}_\lambda \leq \lambda \sum_{x,u} P^{*,\lambda}_{U|X} P_X \log\left(\frac{P^{(n+1)}_{U|X}}{P^{(n)}_{U|X}}\right). \tag{25}$$

*Proof:* See Appendix A. ■

*Theorem 1:* Consider $\lambda \in (0.5, 1]$ and $P^{(0)}_{U|X} \in G_{\delta,\lambda}$ for a given value of $\delta$. If the optimal solution $P^{*,\lambda}_{U|X}$ lies in $H_{\delta,\lambda}(P^{(0)}_{U|X})$ and the function $f(\lambda, P_{U|X})$ is concave in $H_{\delta,\lambda}(P^{(0)}_{U|X})$ and $P^{(0)}_{U|X}$ is such that $|\text{supp}(P^{(0)}_{U|X})| = |\mathcal{U}|$, then:

1) Convergence of $F_\lambda$: $\lim_{n\to\infty} F^{(n)}_\lambda = V_\lambda$;
2) Convergence of $P^{(n)}_{U|X}$: $\lim_{n\to\infty} P^{(n)}_{U|X} = P^{*,\lambda}_{U|X}$.

*Proof:*

[1] It is easy to show that we can always find a value of $\delta$ such that this condition is satisfied.

1) For any integer $N \geq 1$, from Lemma 4 we can bound:

$$\sum_{n=0}^{N-1} V_\lambda - F_\lambda^{(n+1)} \leq \lambda \sum_{x,u} P^{*,\lambda}_{U|X} P_X \log\left(\frac{P^{(N)}_{U|X}}{P^{(0)}_{U|X}}\right) \tag{26}$$

$$= \lambda\left(\mathbb{E}_X\left[\mathcal{D}(P^{*,\lambda}_{U|X} \| P^{(0)}_{U|X})\right] - \mathbb{E}_X\left[\mathcal{D}(P^{*,\lambda}_{U|X} \| P^{(N)}_{U|X})\right]\right) \tag{27}$$

$$\leq \lambda \mathbb{E}_X\left[\mathcal{D}(P^{*,\lambda}_{U|X} \| P^{(0)}_{U|X})\right], \tag{28}$$

where the last term is finite because $|\text{supp}(P^{(0)}_{U|X})| = |\mathcal{U}|$. From claim 2) in Lemma 2 we have: $V_\lambda \geq F_\lambda^{(n+1)}$, and $F_\lambda^{(n)}$ is non-decreasing in $n$. Thus, for $N \to \infty$, the series converges and $F_\lambda^{(n+1)} \overset{n\to\infty}{\longrightarrow} V_\lambda$.

2) From Lemma 2, $F_\lambda^{(n+1)} \geq f(\lambda, P^{(n)}_{U|X}) \geq F_\lambda^{(n)}$, so using the previous claim $f(\lambda, P^{(n)}_{U|X}) \overset{n\to\infty}{\longrightarrow} V_\lambda = f(\lambda, P^{*,\lambda}_{U|X})$. From Lemma 3 we have that $P^{(n)}_{U|X} \in H_{\delta,\lambda}(P^{(0)}_{U|X})$ for all $n$. As $f(\lambda, P_{U|X})$ is concave in $H_{\delta,\lambda}(P^{(0)}_{U|X})$ and its optimal point $P^{*,\lambda}_{U|X}$ is unique, it is easy to check that $P^{(n)}_{U|X} \overset{n\to\infty}{\longrightarrow} P^{*,\lambda}_{U|X}$.

■

As $f(\lambda, P_{U|X})$ is not globally concave, the $\delta$-superlevel set $G_{\delta,\lambda}$ is not convex and it is not necessarily connected. By Lemma 3, the algorithm proposed stays in the path-connected component $H_{\delta,\lambda}(P^{(0)}_{U|X})$ which is determined by the initial condition. If the the optimal point $P^{*,\lambda}_{U|X}$ is contained in the right path-connected component $H_{\delta,\lambda}(P^{(0)}_{U|X})$ and $f(\lambda, P_{U|X})$ is locally concave around the optimal point[2] $P^{*,\lambda}_{U|X}$, the algorithm will converge to the global optimum. However, to avoid convergence to a local maximum located in a wrong path-connected component $G_{\delta,\lambda}$, a simple solution in practice is to run the algorithm from a few different and well-separated initial conditions and keep the one that provides the largest value of $F$ after stopping condition is met. It is worth to mention that this convergence analysis will be valid for other situations where a BA algorithm is used, e.g., for convex and non-convex problems as the standard IB.

### C. Optimal solution and stopping condition

We now consider some properties of the optimal solution $P^*_{U|X}$ and the stopping condition for the proposed algorithm that can be obtained from them. Starting by the next lemma:

*Lemma 5:* Consider $\lambda \in (0.5, 1]$. If $f(\lambda, P_{U|X})$ is concave in a vicinity of the optimal solution $P^{*,\lambda}_{U|X}$, we have

$$\begin{array}{ll} \alpha^*(\lambda, u) = V_\lambda, & \text{for } u \in \mathcal{U} \text{ such that } P^{*,\lambda}_{U|X} > 0 \text{ and } \forall x \in \mathcal{X} \\ \alpha^*(\lambda, u) \leq V_\lambda, & \text{otherwise,} \end{array} \tag{29}$$

where

$$\alpha^*(\lambda, u) := (2\lambda - 1) \sum_{(x,z)\in\mathcal{X}\times\mathcal{Z}} P_{X,Z} \log\left(\frac{Q^{*,\lambda}_{X|Z,U}}{P_{X|Z}}\right)$$
$$+ \lambda \sum_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} P_{X,Y,Z} \log\left(\frac{Q^{*,\lambda}_{U|Y,Z}}{P^{*,\lambda}_{U|X}}\right). \tag{30}$$

*Proof:* See Appendix B. ■

It is interesting to observe that the optimal solution $P^{*,\lambda}_{U|X}$ is such that for each value of $u \in \mathcal{U}$ where $P^{*,\lambda}_{U|X} > 0$ the value of $\alpha^*(\lambda, u)$ is constant and equal to the maximum value $V_\lambda$. Similar results are obtained for the optimum solutions for the capacity of a discrete memoryless channel and the rate-distortion function for a discrete memoryless source [36]. In particular, we have:

$$V_\lambda = \max_{u\in\mathcal{U}} \alpha^*(\lambda, u). \tag{31}$$

From these results, we can consider the quantity:

$$I_\lambda^{(n+1)} := \max_{u\in\mathcal{U}} (2\lambda - 1) \sum_{(x,z)\in\mathcal{X}\times\mathcal{Z}} P_{X,Z} \log\left(\frac{Q^{(n+1)}_{X|Z,U}}{P_{X|Z}}\right)$$
$$+ \lambda \sum_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} P_{X,Y,Z} \log\left(\frac{Q^{(n+1)}_{U|Y,Z}}{P^{(n)}_{U|X}}\right). \tag{32}$$

It is clear from (24) that for $\lambda \in (0.5, 1]$, $I_\lambda^{(n+1)} \geq V_\lambda$. This suggests that a stopping condition specially matched to the optimal value $P^{*,\lambda}_{U|X}$ could be implemented by checking the condition: $I_\lambda^{(n)} - F_\lambda^{(n)} \leq \epsilon$ for a sufficiently small $\epsilon > 0$.

## IV. Numerical Evaluation

In this section, we apply the proposed algorithm to different application problems.

### A. Example of computation of a relevance-rate region

According to the discussion presented in Section II, although region $\mathcal{R}$ is convex, the problem of obtaining its upper-boundary (or the relevance-rate function defined in expression (8)) is not a convex one. For this reason, only a small number of cases can be solved in closed form. One of them is the double binary source problem with binary side information at the decoder which was completed solved in [11]. In this problem, the pmf $P_{X,Y,Z}$ is a probability measure corresponding to a source $(X, Y, Z)$ such that $Y -\!\!\circ\!\!- X -\!\!\circ\!\!- Z$ and $(X, Z)$ form a doubly symmetric binary source with crossover probability $p$, and $(X, Y)$ forms a doubly symmetric binary source with crossover probability $p$. It was shown in [11] that the relevance-rate region is given by the convex hull [32] of the following region:

$$\mathcal{R}_b := \big\{(R, \mu) : R \geq I(X; U|Z),\ \mu \leq I(Y; U|Z),$$
$$\text{with } P_{U|X} = \text{BSC}(r) \quad \forall r \in [0, 0.5]\big\}, \tag{33}$$

where BSC$(r)$ denotes a *binary symmetric channel* with crossover probability $r$. It is clear that the algorithm presented in Section III allows the computation of all relevance-rate pairs in $\mathcal{R}$ for an arbitrary pmf $P_{X,Y,Z}$. This can be easily done by running the algorithm for a sufficient dense grid of points $\lambda \in (0.5, 1]$ for the desired pmf $P_{X,Y,Z}$. In order to test the suitability of the algorithm for this task we used it

[2] This is something reasonable to expect when the optimal point is an interior one because of the smoothness of $f(\lambda, P_{U|X})$.

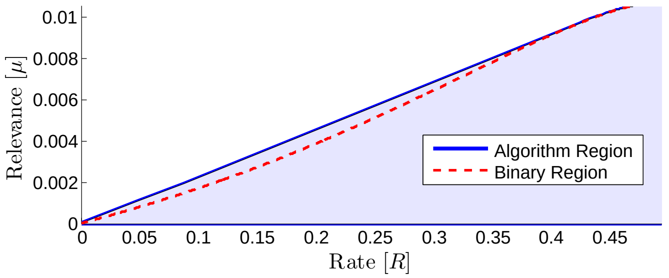

Fig. 4. Rate-relevance region corresponding to a double symmetric binary source.

with the source $(X, Y, Z)$ described above setting parameters $p = 0.1$ and $q = 0.4$. In Fig. 4, we show the region obtained by our algorithm and the upper-boundary of region $\mathcal{R}_b$. We can observe that the region obtained by the algorithm coincides with the convex hull of $\mathcal{R}_b$.

### B. Compression-based regularization learning

In the previous sections, we have shown that the problem of maximizing the relevance $I(U;Y|Z)$ subject to a mutual-information constraint $I(U;X|Z) \leq R$ is equivalent to that of maximizing $f(\lambda, P_{U|X})$ which introduces the penalization term: $(1-\lambda)I(U;X|Z)$. We now show that this constraint can act as a regularization when applied to situations where the joint statistics controlling the observations $P_{X,Y,Z}$ is not known but it is estimated from training samples. Indeed, Shamir *et al.* [37] have already showed evidence that this term can help to prevent "overfitting" and this idea was also exploited in [10], [38] to justify some features of deep learning algorithms. It should mentioned that these analysis were performed for the classical IB method without the presence of side information. The results in [37] can be extended as follows: For any distribution $P_{X,Y,Z}$, with a probability of at least $1-\delta$ over the draw of the sample of size $n$ from $P_{X,Y,Z}$, we have that for any $P_{U|X}$ simultaneously,

$$|I(Y;U|Z) - \hat{I}(Y;U|Z)| \leq \\ K_\delta \frac{\log(n)}{\sqrt{n}} \sqrt{I(X;U|Z)} + O\left(\frac{1}{\sqrt{n}}\right) \quad (34)$$

where $\hat{I}(Y;U|Z)$ is computed using the empirical distribution $\hat{P}_{X,Y,Z}$ based on the $n$ training labeled examples. This result shed some light in how the achievable performance under the empirical distribution approaches the true data distribution one, as a function of the complexity rate $I(X;U|Z)$ showing that looking for the reduction of such term could be beneficial when working this the empirical distribution. Based on these results more precise statements on the generalization learning performance could be obtained assuming a given deviation (which could depend on $n$) between the true distribution $P_{X,Y,Z}$ and the empirical one $\hat{P}_{X,Y,Z}$ and providing a characterization of the achievable performance in terms of rate and relevance when the empirical distribution is used. We defer this study to a future work. In this section, we provide numerical evidence that the desired regularization effects hold in our MTL setup.

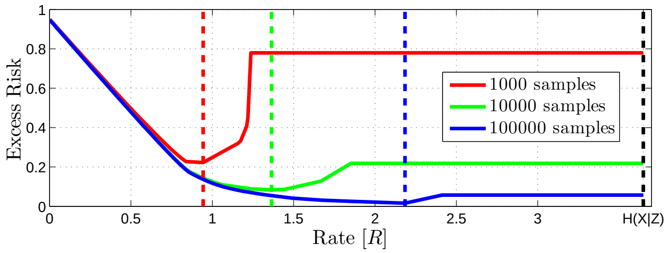

Fig. 5. Excess risk (35) as a function of the information rate.

Consider a multi-task supervised classification problem. Finding the optimal encoder that minimize (3) requires knowledge of the underlying distribution $P_{X,Y,Z}$. From a practical perspective, as the input to the proposed algorithm, we will use the data sampling distribution $\hat{P}_{X,Y,Z}$ based on $n$ training labeled examples. By introducing the rate constraint (or penalization), the optimization problem is reduced to optimizing $L(R, \hat{P}_{X,Y,Z})$ in (8) from which the resulting encoder $\hat{P}^{*,\lambda}_{U|X}$ is derived while the decoder $\hat{P}^{*,\lambda}_{\hat{Y}|U,Z}$ follows from expression (5). The measure of merit will be the *Excess-risk*:

$$\text{Excess-risk} := \text{Risk}\big(\hat{P}^{*,\lambda}_{U|X}, \hat{P}^{*,\lambda}_{\hat{Y}|U,Z}\big) - H(Y|X,Z) \quad (35)$$

that is the difference between the minimum Bayesian risk $H(Y|U,Z)$ and the risk induced from the suboptimal encoder $\hat{P}^{*,\lambda}_{U|X}$ obtained by optimizing w.r.t the sample distribution $\hat{P}_{X,Y,Z}$ subject to the rate constraint.

Experiments will be performed by using synthetic data with alphabets $|\mathcal{X}| = 128$, $|\mathcal{Y}| = 4$, $|\mathcal{Z}| = 2$. The random variable $Z$ is assumed to follow a *Bernoulli* distribution with random parameter $p \in [0,1]$ while the joint distribution $P_{X,Y|Z=z}$ is defined as a *Restrict Boltzmann Machine* (see [1] for further details) with parameters randomly drawn for each $z \in \mathcal{Z}$. The assumed probability $P_Z$ allows for tuning of the relative importance associated to each of the tasks.

In Fig. 5, we plot the excess risk curve as a function of the rate constraint for different size of training samples. With dash lines we denoted the rate for which the excess risk achieves its minimum. When the number of training samples increases the optimal rate $R$ approaches its maximum possible value: $H(X|Z)$ (dashed in black). We emphasize that for every curve there exists a different limiting rate $R_{\text{lim}}$, such that for each $R \geq R_{\text{lim}}$, the excess-risk remains constant with value. It is not difficult to check that $R_{\text{lim}} = \hat{H}(X|Z)$. Furthermore, for every size of training samples, there is an optimal value of $R_{\text{opt}}$ which provides the lowest excess-risk in (35). In a sense, this is indicating that the rate $R$ can be interpreted as an effective regularization term and thus, it can provide robustness for learning in practical scenarios in which the true input distribution is not known and the empirical data distribution is used. It is worth to mention that when more data is available then the optimal value of the regularizing rate $R$ becomes less critical. Of course, this fact was expected since as the amount of training data increases the empirical distribution approaches the true data-generating distribution.

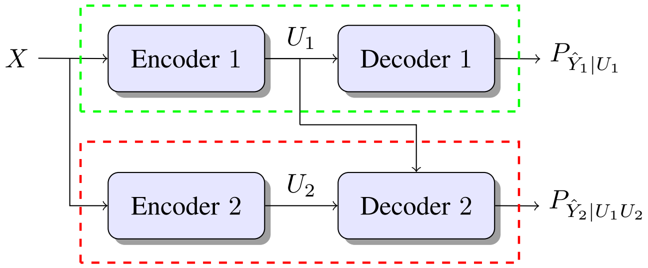

Fig. 6. Hierarchical text categorization and MTL.

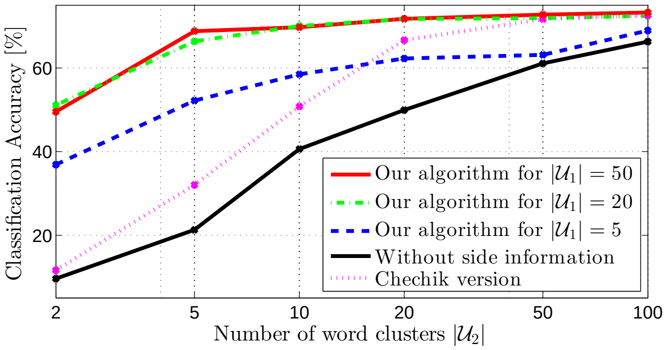

Fig. 7. Classification Accuracy in the hierarchical text categorization problem.

### C. Hierarchical text categorization and distributional word clusters

The high dimensionality of texts can become a severe deterrent in applying complex learners like support vector machines (SVM) [1] to the task of text classification. Word clustering is a powerful alternative to feature selection for reducing the dimensionality of text [6], [39]. This issue can be alleviated by intelligently grouping different classes in disjoint sub-categories. In this way, a first classification problem can be set over the generated sub-categories and the information extracted can be used in a second classification problem to discriminate better between classes. This is the case in hierarchical text classification [17], [40]. We approach this problem based on the scheme of Fig. 6. Consider a document $d$ consisting of different words $X$. We want to estimate the class $Y_2$ to which the document belongs by using information related to a sub-category $Y_1$ (typically related to the text topic) to which the same document also belongs. To this end, assume a pair of encoder 1-decoder 1 infers the document sub-category $\hat{Y}_1$ by using our algorithm without side information (i.e. $Z$ is a degenerate RV) and with input $P_{X,Y_1}$. This is clearly a standard classification problem where $U_1$ is the feature that encoder 1 extracts from $X$. Encoder 2-decoder 2 pair generates the final classification in $\hat{Y}_2$ by using the algorithm with input $P_{X,Y_1,U_1}$. $U_1$ can be considered as side information available at decoder 2. This problem can be interpreted as a MTL problem where the different classification tasks to be inferred by decoder 2 are induced by the features extracted from encoder 1.

Assume a training set consisting of documents belonging to $|\mathcal{Y}_2|$ classes, which has $|\mathcal{X}|$ different words. The distribution $P_{Y_1|Y_2}$ is known because the sub-category $Y_1$ is a deterministic function of the more refined class $Y_2$ (i.e. $Y_1 = h(Y_2)$). The class priors $P_{Y_2}$ are replaced by the empirical distribution and the words distribution conditional to the class, $P_{X|Y_2}$ is estimated using Laplace rule of succession [1]. Imposing the Markov chain $U_1 -\!\!\circ\!\!- X -\!\!\circ\!\!- Y_2$, the resulting joint pmfs are given by $P_{X,Y_2,U_1} = P_{U_1|X}P_{X|Y_2}P_{Y_2}$ and

$$P_{X,Y_1} = \sum_{y_2 \in \mathcal{Y}_2} P_{Y_1|Y_2}P_{X|Y_2}P_{Y_2}. \tag{36}$$

Once pmfs $P_{U_1|X}$ and $P_{U_2|X}$ are calculated using the proposed algorithm, we estimate the class of the document $\hat{y}_2(d)$. Assuming a generative multinomial model, and conditional independence between clusters, the maximum a posteriori probability, which computes the most probable class for a document $d$, is given by (see [39] for details):

$$\begin{aligned} \hat{y}_2(d) &= \arg\max_{y \in \mathcal{Y}_2} P_{Y_2} \prod_{u_1,u_2} \left(P_{U_1,U_2|Y_2}\right)^{n(u_1,u_2,d)} \qquad (37) \\ &= \arg\max_{y \in \mathcal{Y}_2} \log\left(P_{Y_2}\right) + \sum_{u_1,u_2} n(u_1,u_2,d) \log P_{U_1,U_2|Y_2}, \end{aligned}$$

where

$$P_{U_1,U_2|Y_2} = \frac{\sum_x P_{X,Y_2,U_1}P_{U_2|X}}{P_{Y_2}}, \tag{38}$$

and $n(u_1,u_2,d)$ is the number of jointly occurrences of clusters $(u_1,u_2)$ in the document $d$ computed with:

$$u_i(x) := \arg\max_u P_{U_i|X}(u|x), \qquad i \in \{1,2\}. \tag{39}$$

We test the above proposed classification procedure on the 20 Newsgroups (20Ng) dataset [41]. This contains 11269 documents for training and 7505 for testing evenly divided among 20 UseNet Discussion groups or classes. Each newsgroup represents one class in the classification task. The train dataset had 53975 different words. The 20 Newsgroups correspond to 6 topics. The sub-category $Y_1$ represents the topic among 6 possibilities and the refined classification $Y_2$ is the class among 20 possibilities.

In Fig. 7, our algorithm performance ($\lambda = 0.99$) versus $|\mathcal{U}_2|$ is compared with the algorithm without side information (which is a single-task setup) and the one proposed in [17]. It is interesting to mention that the single-task setting and the one in [17] can be covered using the proposed algorithm. In particular, for the single-task setup, we estimate the final class with $P_{X,Y_2}$ as the input of the proposed algorithm ($Z$ is a degenerate RV). Our setting and the one in [17] show an improvement with respect to the single task setup without side information. This suggests that exploiting the common features in MTL may be advantageous. With $|\mathcal{U}_1| = 20$ and $|\mathcal{U}_2| = 5$, our method achieves 66.36% of accuracy. For which we exploit the additional information in a structured manner to show an improvement with respect to the other proposals.

A variation of this application is Distributional Word Clusters (DWC) introduced in [42] where the authors develop an IB algorithm (without side information) for distributions of verb-object pairs. Clusters derived by their algorithm seem in many cases semantically significant, and it can be used to study the linguistic structure of a language. The DWC is also

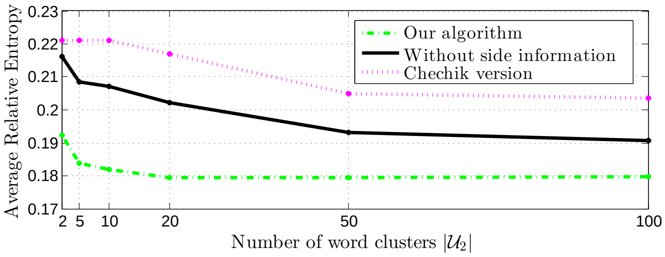

Fig. 8. Average relative entropy in the DWC problem.

used to relevant text categorization features extraction [43]. We consider the same setting in Fig. 6 and study the average relative entropy metric given by:

$$\frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}}\mathcal{D}\left(P_{Y_2|X}(\cdot|x)\big\|P_{\hat{Y}_2|X,U_1}(\cdot|x,u_1(x))\right) \tag{40}$$

where $|\mathcal{U}_1| = 20$,

$$P_{Y_2|X} = \frac{P_{X|Y_2}P_{Y_2}}{\sum_x P_{X|Y_2}P_{Y_2}} \tag{41}$$

is computed on the test set and $P_{\hat{Y}_2|X,U_1}$ is defined by

$$P_{\hat{Y}_2|X,U_1} = \sum_{u_2\in\mathcal{U}_2} P_{U_2|X}P_{Y_2|U_1,U_2}, \tag{42}$$

$$P_{Y_2|U_1,U_2} = \frac{\sum_x P_{U_2|X}P_{X,Y_2,U_1}}{\sum_x P_{U_2|X}P_{U_1|X}P_X}. \tag{43}$$

These measurements of model quality were introduced in [42]. While the setting in [17] has worse performance than the IB without side information, our proposed framework introduces some advantages showing an improved use of the side information to enforce similarity between $P_{Y_2|X}$ and less complex models (fewer number of clusters). In other words, the MTL information-theoretic model behaves in the expected sense with respect to generalization and model complexity.

*Remark 1:* There exists a strong relationship between our objective and the one in [17] when we redefine the tasks as the classification within the different sub-categories. In this case, referring to Fig. 2 we consider $Y = Y_2$ and the side information as a deterministic function of the task $Z = g(Y)$ (every class in the same sub-category belongs to the same $Z$). Our objective $f(\lambda, P_{U|X})$ can be written as:

$$\begin{aligned} f(\lambda, P_{U|X}) &= \lambda I(Y;U|g(Y)) - \bar{\lambda} I(X;U|g(Y)) \qquad (44)\\ &= \lambda\left[I(Y;U) - I(g(Y);U)\right] - \bar{\lambda}\left[I(X;U) - I(g(Y);U)\right]\\ &= \lambda I(Y;U) - (2\lambda-1)I(g(Y);U) - \bar{\lambda}I(X;U) \end{aligned}$$

where $\bar{\lambda} = 1-\lambda$ and $f(\lambda, P_{U|X})$ depends on the source via the marginal distributions: $P_{X,Y}$ and $P_{X,Z}$. This expression is the cost function proposed in [17] with $\beta = \frac{\lambda}{1-\lambda}$ and $\gamma = \frac{2\lambda-1}{\lambda}$.

## V. Conclusions

From information-theoretic methods, we have investigated the supervised learning framework of MTL in which an encoder builds a common representation intended to several related tasks. The obtained formulation can be seen as the problem of obtaining the rate-relevance region of an information bottleneck problem with side information only at the decoder. With this interpretation, we derived an iterative learning algorithm from the principle of compression-based regularization as a natural safeguard against overfitting. Numerical evidence showed that there exists an optimal compression rate minimizing the *excess risk* according to the amount of available training data. Indeed, this rate increases with the size of the training set. Applications to hierarchical text categorization were also considered.

It should be mentioned that several open questions remain regarding the statistical regularization properties of building compact representations of data. It is clear that both further theoretical and practical studies are required. Applications of our algorithm to other MTL setups, besides the hierarchical text categorization one, should also deserve additional efforts.

One of the most important working hypotheses to be modified would be the consideration of continuous sources. Although the analysis would be harder we think that the results could be of interest for other learning applications were the examples are of continuous nature (e.g. images, speech signals, etc). Studies regarding variations on the learning cost function used (e.g. minimizing the maximum log-loss over the different tasks instead of the average one) also deserve further study.

## Appendix

### Appendix A: Proof of Lemma 4

In order to show Lemma 4, we will need the following auxiliary result:

*Lemma 6:* Let $\lambda \in (0.5, 1]$ and $T$ be a convex set of conditional distributions $P_{U|X}$ such that the function $f(\lambda, P_{U|X})$ is concave in the domain $T$. Then, $\mathcal{L}[P_a, P_b] \geq 0$ for any $P_a, P_b \in T$, where

$$\begin{aligned} \mathcal{L}[P_a, P_b] = \sum_{(x,y,z,u)} & P_a P_{X,Y,Z}\left[\lambda\mathcal{D}\left(P_a\|P_b\right)\right.\\ & - \lambda\mathcal{D}\left(Q_{U|Y,Z}[P_a]\|Q_{U|Y,Z}[P_b]\right)\\ & \left. - (2\lambda-1)\mathcal{D}\left(Q_{X|U,Z}[P_a]\|Q_{X|U,Z}[P_b]\right)\right], \end{aligned} \tag{45}$$

and we have defined, for $i = \{a, b\}$:

$$Q_{U|Y,Z}[P_i] = \sum_{x\in\mathcal{X}} P_i P_{X|Y,Z}, \quad Q_{X|Z,U}[P_i] = \frac{P_i P_{X|Z}}{\sum_x P_i P_{X|Z}}. \tag{46}$$

*Proof:* We start calculating $\frac{\partial f(\lambda, P_{U|X})}{\partial P_{U|X}}$. For $(u, x)$ such that $P_{U|X} = 0$ the derivative is zero. For $(u, x)$ such that $P_{U|X} > 0$, we use the identity: $[f(x)\log(f(x))]' = f'(x)\log(ef(x))$ and obtain:

$$\begin{aligned} \frac{\partial f(\lambda, P_{U|X})}{\partial P_{U|X}} &= (2\lambda-1)P_X\log\left(eP_{U|X}\right)\\ &- (2\lambda-1)\sum_z P_{X,Z}\log\left(eP_{U|Z}\right)\\ &- \lambda\, P_X\log\left(eP_{U|X}\right) + \lambda\sum_{y,z} P_{X,Y,Z}\log\left(eP_{U|Y,Z}\right) \qquad (47)\\ &= (2\lambda-1)\sum_z P_{X,Z}\log\left(\frac{P_{U|X}}{P_{U|Z}}\right) \end{aligned}$$

$$-\lambda\sum_{y,z}P_{X,Y,Z}\log\left(\frac{P_{U|X}}{P_{U|Y,Z}}\right) \tag{48}$$

$$\begin{aligned}&=(2\lambda-1)\sum_{z}P_{X,Z}\log\left(\frac{P_{X|Z,U}}{P_{X|Z}}\right)\\ &\quad-\lambda\sum_{y,z}P_{X,Y,Z}\log\left(\frac{P_{U|X}}{P_{U|Y,Z}}\right).\end{aligned} \tag{49}$$

Note that

$$f(\lambda,P_{U|X})=\sum_{x,u}P_{U|X}\frac{\partial f(\lambda,P_{U|X})}{\partial P_{U|X}}. \tag{50}$$

Then,

$$\sum_{x,u}P_b\left.\frac{\partial f(\lambda,P_{U|X})}{\partial P_{U|X}}\right|_{P_b}=f(\lambda,P_b). \tag{51}$$

Now, let us consider:

$$\begin{aligned}&\sum_{x,u}P_a\left.\frac{\partial f(\lambda,P_{U|X})}{\partial P_{U|X}}\right|_{P_b}\\ &=f(\lambda,P_a)-(2\lambda-1)\sum_{x,z,u}P_aP_{X,Z}\log\left(\frac{Q_{X|Z,U}[P_a]}{Q_{X|Z,U}[P_b]}\right)\\ &+\lambda\sum_{x,u}P_aP_X\log\left(\frac{P_a}{P_b}\right)\\ &-\lambda\sum_{x,y,z,u}P_aP_{X,Y,Z}\log\left(\frac{Q_{U|Y,Z}[P_a]}{Q_{U|Y,Z}[P_b]}\right)\end{aligned} \tag{52}$$

$$\begin{aligned}&=f(\lambda,P_a)\\ &+\sum_{x,y,z,u}P_aP_{X,Y,Z}\left[-(2\lambda-1)\mathcal{D}\left(Q_{X|Z,U}[P_a]\|Q_{X|Z,U}[P_b]\right)\right.\\ &\left.+\lambda\mathcal{D}\left(P_a\|P_b\right)-\lambda\mathcal{D}\left(Q_{U|Y,Z}[P_a]\|Q_{U|Y,Z}[P_b]\right)\right]\end{aligned} \tag{53}$$

$$=f(\lambda,P_a)+\mathcal{L}[P_a,P_b]. \tag{54}$$

Then,

$$\mathcal{L}[P_a,P_b]=\sum_{x,u}P_a\left.\frac{\partial f(\lambda,P_{U|X})}{\partial P_{U|X}}\right|_{P_b}-f(\lambda,P_a) \tag{55}$$

$$=\sum_{x,u}(P_a-P_b)\left.\frac{\partial f(\lambda,P_{U|X})}{\partial P_{U|X}}\right|_{P_b}-f(\lambda,P_a)+f(\lambda,P_b). \tag{56}$$

If $f(\lambda,P_{U|X})$ is concave in $T$, then:

$$f(\lambda,P_a)\le f(\lambda,P_b)+\sum_{x,u}\left.\frac{\partial f(\lambda,P_{U|X})}{\partial P_{U|X}}\right|_{P_b}(P_a-P_b), \tag{57}$$

and thus: $\mathcal{L}[P_a,P_b]\ge f(\lambda,P_a)-f(\lambda,P_a)=0$. ■

Now we can proceed to the proof of Lemma 4. In order to show (24), we define the quantity $\mathcal{B}$:

$$\begin{aligned}\mathcal{B}:=(2\lambda-1)&\sum_{x,z,u}P^{*,\lambda}_{U|X}P_{X,Z}\log\left(\frac{Q^{(n+1)}_{X|Z,U}}{P_{X|Z}}\right)\\ &+\lambda\sum_{x,y,z,u}P^{*,\lambda}_{U|X}P_{X,Y,Z}\log\left(\frac{Q^{(n+1)}_{U|Y,Z}}{P^{(n)}_{U|X}}\right).\end{aligned} \tag{58}$$

Then, we can write:

$$\begin{aligned}V_\lambda=(2\lambda-1)&\sum_{x,z,u}P^{*,\lambda}_{U|X}P_{X,Z}\log\left(\frac{Q^{*}_{X|Z,U}}{P_{X|Z}}\right)\\ &+\lambda\sum_{x,y,z,u}P^{*,\lambda}_{U|X}P_{X,Y,Z}\log\left(\frac{Q^{*}_{U|Y,Z}}{P^{*,\lambda}_{U|X}}\right)\end{aligned} \tag{59}$$

$$\begin{aligned}&=\mathcal{B}+(2\lambda-1)\sum_{x,z,u}P^{*,\lambda}_{U|X}P_{X,Z}\mathcal{D}\left(Q^{*}_{X|Z,U}\|Q^{(n+1)}_{X|Z,U}\right)\\ &\quad+\lambda\sum_{y,z}P_{Y,Z}\mathcal{D}\left(Q^{*}_{U|Y,Z}\|Q^{(n+1)}_{U|Y,Z}\right)\\ &\quad-\lambda\sum_{x}P_X\mathcal{D}\left(P^{*,\lambda}_{U|X}\|P^{(n)}_{U|X}\right)\end{aligned} \tag{60}$$

$$=\mathcal{B}-\mathcal{L}[P^{*,\lambda}_{U|X},P^{(n)}_{U|X}]. \tag{61}$$

Consider an integer $n\ge 1$ and the set $\tilde{G}^{n}_{\delta,\lambda}$ from the proof of Lemma 3. It is known that this set is convex and from its definition should contain $P^{(n)}_{U|X}$ and the optimal solution $P^{*,\lambda}_{U|X}$. As the function $f(\lambda,P_{U|X})$ is concave in $H_{\delta,\lambda}(P^{(0)}_{U|X})$ and $\tilde{G}^{n}_{\delta,\lambda}\subseteq H_{\delta,\lambda}(P^{(0)}_{U|X})$, we can apply Lemma 6 to $\mathcal{L}[P^{*,\lambda}_{U|X},P^{(n)}_{U|X}]$ and conclude that $V_\lambda\le\mathcal{B}$. We also define:

$$\begin{aligned}\gamma^{(n+1)}_{U|X}=\exp\Bigg\{&\frac{2\lambda-1}{\lambda}\sum_{z}P_{Z|X}\log(Q^{(n+1)}_{X|Z,U})\\ &+\sum_{y,z}P_{Y|X,Z}P_{Z|X}\log(Q^{(n+1)}_{U|Y,Z})\Bigg\},\end{aligned} \tag{62}$$

from which it is clear that $P_{U|X}\propto\gamma^{(n+1)}_{U|X}$. It is not hard to see that:

$$\mathcal{B}=\lambda\sum_{x,u}P^{*,\lambda}_{U|X}P_X\log\left(\frac{\gamma^{(n+1)}_{U|X}}{P^{(n)}_{U|X}}\right)+(2\lambda-1)H(X|Z). \tag{63}$$

On the other hand, $F^{(n+1)}_{\lambda}$ can be written as:

$$\begin{aligned}F^{(n+1)}_{\lambda}=(2\lambda-1)&\sum_{x,z,u}P^{(n+1)}_{U|X}P_{X,Z}\log\left(\frac{Q^{(n+1)}_{X|Z,U}}{P_{X|Z}}\right)\\ &-\lambda\sum_{x,y,z,u}P^{(n+1)}_{U|X}P_{X,Y,Z}\log\left(\frac{\gamma^{(n+1)}_{U|X}}{Q^{(n+1)}_{U|Y,Z}\sum_{u'}\gamma^{(n+1)}_{U'|X}}\right)\end{aligned} \tag{64}$$

$$\begin{aligned}&=(2\lambda-1)\sum_{x,z,u}P^{(n+1)}_{U|X}P_{X,Z}\log\left(\frac{Q^{(n+1)}_{X|Z,U}}{P_{X|Z}}\right)\\ &\quad+\lambda\sum_{x,y,z,u}P^{(n+1)}_{U|X}P_{X,Y,Z}\log\left(Q^{(n+1)}_{U|Y,Z}\sum_{u'}\gamma^{(n+1)}_{U'|X}\right)\\ &\quad-(2\lambda-1)\sum_{x,z,u}P^{(n+1)}_{U|X}P_{X,Z}\log(Q^{(n+1)}_{X|Z,U})\\ &\quad-\lambda\sum_{x,y,z,u}P^{(n+1)}_{U|X}P_{X,Y,Z}\log(Q^{(n+1)}_{U|Y,Z})\end{aligned} \tag{65}$$

$$=\lambda\sum_{x}P_X\log\left(\sum_{u'}\gamma^{(n+1)}_{U'|X}\right)+(2\lambda-1)H(X|Z). \tag{66}$$

Finally,

$$V_\lambda - F_\lambda^{(n+1)} \leq \mathcal{B} - \lambda \sum_x P_X \log \left( \sum_{u'} \gamma_{U'|X}^{(n+1)} \right) \\ - (2\lambda - 1) H(X|Z) \qquad (67)$$

$$= \lambda \sum_{x,u} P_{U|X}^{*,\lambda} P_X \log \left( \frac{P_{U|X}^{(n+1)}}{P_{U|X}^{(n)}} \right). \qquad (68)$$

## Appendix B: Proof of Lemma 5

The proofs is along the lines of Karush-Kuhn-Tucker (KKT) conditions [31]. In this case we look for the maximum of $f(\lambda, P_{U|X})$ subject to $\sum_u P_{U|X} = 1$ for all $x \in \mathcal{X}$ and $P_{U|X} \geq 0$ for all $(u,x) \in \mathcal{U} \times \mathcal{X}$. Provided that $f(\lambda, P_{U|X})$ is concave in a vicinity of $P_{U|X}^{*,\lambda}$ a necessary condition for the local optimality of $P_{U|X}^{*,\lambda}$ is the existence of values $\phi_{x,u}, \kappa_x$ such that

1) $\frac{\partial f(\lambda, P_{U|X}^{*,\lambda})}{\partial P_{U|X}} = \kappa_x - \phi_{x,u}$ for all $(u,x) \in \mathcal{U} \times \mathcal{X}$,
2) $\sum_u P_{U|X}^{*,\lambda} = 1$ for all $x \in \mathcal{X}$,
3) $P_{U|X}^{*,\lambda} \geq 0$ for all $(u,x) \in \mathcal{U} \times \mathcal{X}$,
4) $\phi_{x,u} \geq 0$ for all $(u,x) \in \mathcal{U} \times \mathcal{X}$,
5) $\phi_{x,u} P_{U|X}^{*,\lambda} = 0$ for all $(u,x) \in \mathcal{U} \times \mathcal{X}$.

From conditions 1) and 4), we obtain for all $(u,x) \in \mathcal{U} \times \mathcal{X}$:

$$\kappa_x \geq \frac{\partial f(\lambda, P_{U|X}^{*,\lambda})}{\partial P_{U|X}}. \qquad (69)$$

From condition 5), we observe that equality is achieved for all $(u,x) \in \mathcal{U} \times \mathcal{X}$ such that $P_{U|X}^{*,\lambda} > 0$. From (49) we have:

$$\frac{\partial f(\lambda, P_{U|X}^{*,\lambda})}{\partial P_{U|X}} = (2\lambda - 1) \sum_z P_{X,Z} \log \left( \frac{Q_{X|Z,U}^*}{P_{X|Z}} \right) \\ - \lambda \sum_{y,z} P_{X,Y,Z} \log \left( \frac{P_{U|X}^{*,\lambda}}{Q_{U|Y,Z}^*} \right). \qquad (70)$$

Combining these two last equations and summing over all $x \in \mathcal{X}$, we have:

$$\sum_x \kappa_x \geq (2\lambda - 1) \sum_{x,z} P_{X,Z} \log \left( \frac{Q_{X|Z,U}^*}{P_{X|Z}} \right) \\ - \lambda \sum_{x,y,z} P_{X,Y,Z} \log \left( \frac{P_{U|X}^{*,\lambda}}{Q_{U|Y,Z}^*} \right) \qquad (71)$$

$$= \alpha^*(\lambda, u). \qquad (72)$$

In a similar manner, from conditions 1) and 5) and using Eq. (50), we can write:

$$V_\lambda = f(\lambda, P_{U|X}^{*,\lambda}) = \sum_{x,u} P_{U|X}^{*,\lambda} \frac{\partial f(\lambda, P_{U|X}^{*,\lambda})}{\partial P_{U|X}} \qquad (73)$$

$$= \sum_{x,u} P_{U|X}^{*,\lambda} (\kappa_x - \phi_{x,u}) \qquad (74)$$

$$= \sum_x \kappa_x. \qquad (75)$$

It is straightforward to check that $V_\lambda \geq \alpha^*(\lambda, u)$ for all $u \in \mathcal{U}$. Finally, from condition 5) and by similar arguments, it is easy to see that $V_\lambda = \alpha^*(\lambda, u)$ $\forall$ $(u,x) \in \mathcal{U} \times \mathcal{X}$ s.t. $P_{U|X}^{*,\lambda} > 0$.

## References

[1] K. P. Murphy, *Machine learning: a probabilistic perspective*, Cambridge, MA, 2012.

[2] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, 1999, pp. 368–377.

[3] R. Dobrushin and B. Tsybakov, "Information transmission with additional noise," *IRE Transactions on Information Theory*, vol. 8, no. 5, pp. 293–304, September 1962.

[4] K. Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2210–2239, Nov 1998.

[5] H. Witsenhausen and A. Wyner, "A conditional entropy bound for a pair of discrete random variables," *Information Theory, IEEE Transactions on*, vol. 21, no. 5, pp. 493–501, 1975.

[6] N. Slonim and N. Tishby, "The power of word clusters for text classification," in *23rd European Colloquium on Information Retrieval Research (ECIR)*, 2001, pp. 1–12.

[7] N. Slonim, R. Somerville, N. Tishby, and O. Lahav, "Objective classification of galaxy spectra using the information bottleneck method," in *Monthly Notes of the Royal Astronomical Society*, vol. 323, 2001, pp. 270–284.

[8] R. M. Hecht, E. Noor, and N. Tishby, "Speaker recognition by gaussian information bottleneck." in *INTERSPEECH*. ISCA, 2009, pp. 1567–1570.

[9] M. Vera, L. R. Vega, and P. Piantanida, "The two-way cooperative information bottleneck," in *IEEE International Symp. on Information Theory, ISIT 2015*, 2015, pp. 2131–2135.

[10] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *2015 IEEE Information Theory Workshop, ITW 2015, Jerusalem, Israel, 2015*, 2015, pp. 1–5.

[11] M. Vera, L. Rey Vega, and P. Piantanida, "Collaborative representation learning," *ArXiv e-prints*, Apr. 2016. [Online]. Available: http://arxiv.org/abs/1604.01433

[12] G. Pichler, P. Piantanida, and G. Matz, "Distributed information-theoretic biclustering," *CoRR*, vol. abs/1602.04605, 2016. [Online]. Available: http://arxiv.org/abs/1602.04605

[13] Q. Yang, P. Piantanida, and D. Gunduz, "The multi-layer information bottleneck problem," in *Information Theory Workshop (ITW), 2017 IEEE, Nov. 6th – 10th*, 2017.

[14] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inform. Theory*, vol. 18, no. 1, pp. 14–20, 1972.

[15] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. 18, no. 4, pp. 460–473, 1972.

[16] F. M. Willems, *Computation Wyner-Ziv rate-distortion function*, ser. Eindhoven University of Technology Research Reports, Jul. 1983.

[17] G. Chechik and N. Tishby, "Extracting relevant structures with side information," in *Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada]*, 2002, pp. 857–864.

[18] G. Kumar and A. Thangaraj, "Computation of secrecy capacity for more-capable channel pairs," in *Information Theory (ISIT), 2008 IEEE International Symposium on*, Toronto, Canada, July 2008.

[19] K. Yasui, T. Suko, and T. Matsushima, "An algorithm for computing the secrecy capacity of broadcast channels with confidential messages," in *Information Theory (ISIT), 2007 IEEE International Symposium on*, Nice, France, June 2007.

[20] K. Yasui and T. Matsushima, "Toward computing the capacity region of degraded broadcast channel," in *Information Theory (ISIT), 2010 IEEE International Symposium on*, Texas, U.S.A, June 2010.

[21] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, Jul 1997. [Online]. Available: https://doi.org/10.1023/A:1007379606734

[22] J. Baxter, "A model of inductive bias learning," *Journal of Artificial Intelligence Research*, vol. 12, pp. 149–198, 2000.

[23] Y. Zhang and Q. Yang, "A Survey on Multi-Task Learning," *arXiv:1707.08114 [cs]*, Jul. 2017, arXiv: 1707.08114. [Online]. Available: http://arxiv.org/abs/1707.08114

[24] R. K. Ando and T. Zhang, "A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data," *Journal of Machine Learning Research*, vol. 6, no. Nov, pp. 1817–1853, 2005.

[25] C. Ciliberto, Y. Mroueh, T. Poggio, and L. Rosasco, "Convex Learning of Multiple Tasks and their Structure," *arXiv:1504.03101 [cs]*, Apr. 2015, arXiv: 1504.03101. [Online]. Available: http://arxiv.org/abs/1504.03101

[26] A. Argyriou, T. Evgeniou, and M. Pontil, “Convex Multi-task Feature Learning,” *Mach. Learn.*, vol. 73, no. 3, pp. 243–272, Dec. 2008.

[27] M. L. Zhang and Z. H. Zhou, “A Review on Multi-Label Learning Algorithms,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.

[28] A. D. Wyner and J. Ziv, “The rate-distortion function for source coding with side information at the decoder,” *IEEE Trans. Inform. Theory*, vol. 22, pp. 1–10, 1976.

[29] A. El Gamal and Y.-H. Kim, *Network Information Theory*. New York, NY, USA: Cambridge University Press, 2012.

[30] M. Li and P. Vitanyi, “An introduction to kolmogorov complexity and its applications: Preface to the first edition,” 1997.

[31] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, USA: Cambridge University Press, 2004.

[32] R. T. Rockafellar, *Convex Analysis*. Princeton University Press, Jun. 1970.

[33] M. Chalk, O. Marre, and G. Tkacik, “Relevant sparse codes with variational information bottleneck,” in *Advances in Neural Information Processing Systems (NISP 2016)*, 2016, pp. 1957–1965.

[34] A. Alemi, I. Fischer, J. Dillon, and K. Murphy, “Deep variational information bottleneck,” in *International Conference of Learning Representation (ICLR 2017)*, 2017.

[35] A. Painsky and N. Tishby, “Gaussian lower bound for the information bottleneck limit,” *CoRR*, vol. abs/1711.02421, 2017.

[36] R. Gallager, *Information Theory and Reliable Communication*. New York, USA: John Wiley & Sons, Inc., 1968.

[37] O. Shamir, S. Sabato, and N. Tishby, “Learning and generalization with the information bottleneck,” *Theor. Comput. Sci.*, vol. 411, no. 29-30, pp. 2696–2711, Jun. 2010. [Online]. Available: http://dx.doi.org/10.1016/j.tcs.2010.04.006

[38] R. Shwartz-Ziv and N. Tishby, “Opening the black box of deep neural networks via information,” *CoRR*, vol. abs/1703.00810, 2017.

[39] I. S. Dhillon, S. Mallela, and R. Kumar, “A divisive information theoretic feature clustering algorithm for text classification,” *J. Mach. Learn. Res.*, vol. 3, pp. 1265–1287, Mar. 2003.

[40] A. Vinokourov and M. Girolami, “A Probabilistic Framework for the Hierarchic Organisation and Classification of Document Collections,” *Journal of Intelligent Information Systems*, vol. 18, no. 2–3, pp. 153–172, Mar. 2002.

[41] K. Lang, “Newsweeder: Learning to filter netnews,” in *in Proceedings of the 12th International Machine Learning Conference (ML95)*, 1995.

[42] F. Pereira, N. Tishby, and L. Lee, “Distributional clustering of english words,” in *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*, ser. ACL ’93, 1993, pp. 183–190.

[43] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter, “Distributional word clusters vs. words for text categorization,” *Journal of Machine Learning Research*, vol. 3, Mar. 2003.