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

Matias Vera, Leonardo Rey Vega, Pablo Piantanida

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 by introducing the noisy lossy source coding paradigm with the log-loss fidelity criterion which provides the fundamental tradeoffs between the *cross-entropy loss* (average risk) and the information rate of the features (model complexity). Our approach allows an information theoretic formulation of the *multi-task learning* (MTL) problem which is a supervised learning framework in which the prediction models for several related tasks are learned jointly from common representations to achieve better generalization performance. Then, we present an iterative algorithm for computing the optimal tradeoffs and its global convergence is proven provided that some conditions hold. An important property of this algorithm is that it provides 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. An application to hierarchical text categorization is also investigated, extending previous works.**

***Index Terms*—Multi-task learning, Rate distortion, Data Compression, Regularization, Risk, Information Bottleneck, 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. The actual goal of learning is neither accurate estimation of model parameters nor compact representation of the data itself; rather, we are interested in the generalization capabilities, i.e., its ability to successfully apply rules extracted from previously seen data to characterize unseen data. 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, 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.

Shannon's seminal work [1] on information compression with a fidelity criterion provides a function for measuring the distortion (or loss) between the original signal and its compressed representation. The rate-distortion function is 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 [2] and references therein). This paper is concerned with an iterative algorithm for computing the rate-distortion of a generalization of Shannon's model, referred to as noisy source coding with the log-loss fidelity and side information, and its applications to multi-task learning.

A. Related work

The noisy source coding problem was first introduced by Dobrushin and Tsybakov [3] with the goal of generating a good description of an observed source Y (at the encoder) in order to minimize its average distortion with respect to its reconstructed version (at the decoder). The main difference with respect to the original Shannon's problem relies on that Y is not observed directly at the encoder. Instead, a noisy version of Y denoted by X is observed and appropriately compressed. More precisely, for a memoryless source with single-letter distribution P_Y observed through a noisy channel with single-input transition probability $P_{X|Y}$, the noisy distortion-rate function under an additive distortion measure is given by

$$L(R, P_{XY}) \coloneqq \min_{P_{U|X}:\, I(U;X)\leq R} \mathbb{E}_{P_{XY}P_{U|X}}[\ell(U,Y)]. \quad (1)$$

Motivated by the fact that it is not always obvious what loss function $\ell(u, y)$ should be used, especially if the data cannot be structured in a metric space (e.g. speech), Tishby *et al.* [4] associated this information-theoretic setup to a learning problem in which the encoder builds a (compressed) feature U by

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 - Universit Paris-Sud, France (e-mail: pablo.piantanida@centralesupelec.fr).

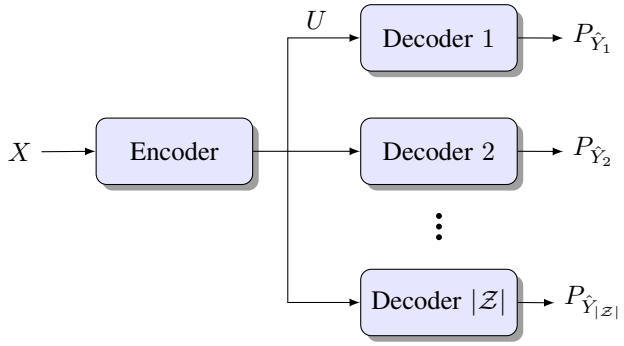

Fig. 1. Multi-task learning problem where X denotes the input data, U is the common feature (data representation) and the corresponding soft-estimates of the multiple labels are denoted by $P_{\hat{Y}_1|U}, \ldots, P_{\hat{Y}_{|\mathcal{Z}|}|U}$.

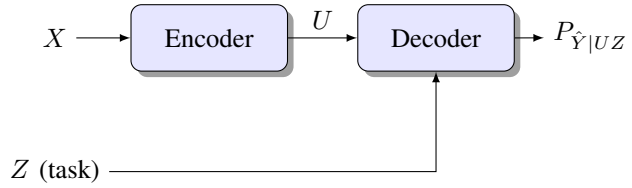

Fig. 2. Multi-task learning as the noisy source coding problem with log-loss fidelity and side information Z (index task variable) at the decoder, referred to as the IB with side information.

extracting from data X information about another variable Y. The idea of the so-called *Information Bottleneck* (IB) method is to identify relevant information from observed samples as being the information that those observations provide about another hidden signal (e.g., the information that face images X provide about the names of the people portrayed Y). To this end, the IB introduces the *log-loss fidelity*:

$$\ell(u, y) \coloneqq -\log P_{Y|U}(y|u), \tag{2}$$

where the *soft-decoder* probability is obtained as:

$$P_{Y|U}(y|u) = \frac{\sum_{x\in\mathcal{X}} P_{U|X}(u|x)P_{XY}(x,y)}{\sum_{x\in\mathcal{X}} P_{U|X}(u|x)P_X(x)}, \tag{3}$$

which is clearly determined by the *soft-encoder* $P_{U|X}$ and the data distribution P_{XY}. The optimal $P^{\star}_{U|X}$ is computed as the solution minimizing (1). This problem can be formulated using duality in optimization theory [5] which leads to:

$$P^{\star}_{U|X} \coloneqq \arg\min_{P_{U|X}} \mathbb{E}_{\hat{P}_{XY}P_{U|X}}[-\log P_{Y|U}(Y|U)] + \beta I(U;X), \tag{4}$$

where the Lagrange multiplier $\beta \geq 0$ is a parameter that controls the tradeoff between compression rate R and the average log-loss. However, the expectation is taken w.r.t. the sampling distribution $\hat{P}_{XY}$ because in real-world problems the true data distribution P_{XY} is not known. Notice that the Lagrange multiplier β can be interpreted as a parametrization of rate R. In this sense, it is clear that there is a family of optimal solutions, one for each R or equivalently β.

An interesting variation is given by the observation that in the above problem the decoder is completely determined by the encoder and the data distribution. We can however consider that both $(P_{U|X}, P_{\hat{Y}|U})$ have to be optimized:

$$\begin{aligned}(P^{\star}_{U|X}, P^{\star}_{\hat{Y}|U}) \;\coloneqq\; & \arg\min_{P_{U|X}, P_{\hat{Y}|U}} \mathbb{E}_{\hat{P}_{XY}P_{U|X}}[-\log P_{\hat{Y}|U}(Y|U)] \\ & + \beta I(U;X),\end{aligned} \tag{5}$$

which can be seen as the optimization of a penalized *cross-entropy* metric[1]. Different from (4), this problem does not lead (at least to our knowledge) to an information theory operational meaning as (4). However, it is easy to check that given an arbitrary encoder $P_{U|X}$, the optimal decoder choice is given by (3). Therefore, expression (5) is –from the point of view of the optimization problem– not more general than (4). For this reason and the connection with noisy source coding with log-loss fidelity, we will be concentrate our efforts in (4).

Witsenhausen and Wyner [7] were the first studying an information-theoretic problem equivalent to (1) and obtained an interesting characterization of its solution and several applications to source coding. Whereas the IB method, in the same way as we presented it above, was introduced in [4] as a rate-distortion problem with an additive fidelity measure. Since then, it was applied to derive several clustering algorithms for a wide variety of applications such as: text classification [8], galaxy spectra classification [9], speaker recognition [10], among others. Further information-theoretic extensions of the IB were recently considered in [11]–[15]. In particular, in some of these works it is shown that the same characterization of the rate and distortion tradeoff is obtained when the distortion metric is not necessarily additive as assumed in (2).

The optimization in (4) was typically addressed by resorting to Blahut–Arimoto (BA) type algorithms. These are often used to refer to a class of algorithms for numerically computing the capacity of a noisy channel and the rate-distortion function for given source. They are iterative algorithms that eventually converge to the optimal solution provided that the optimization problem is convex. The algorithm for the classical rate-distortion problem, i.e., by setting $X = Y$ in (1), was developed independently by Arimoto [16] and Blahut [17]. An extension of this algorithm to the rate-distortion function with side information at the decoder was reported in [18]. Although this algorithm can be applied for optimizing the IB criterion [4], we emphasize that conventional algorithms [16], [17] are only expected to converge to a local minimum since expression (4) leads to a non-convex problem due to the presence of the soft-encoder in the fidelity measure in (2). Chechik *et al.* [19] adapts a BA algorithm to a restricted form of side information without further study the involved algorithm. In a different but related optimization problem, Kumar and Thangaraj [20] adapt the BA algorithm and analysis techniques provided in [21] to a non-convex problem while Yasui and Matsushima [22] extend this work for computing rate regions.

In this paper, we present a novel algorithm for multi-task learning based on the IB paradigm with side information. Multi-task Learning (MTL) [23] is an approach to inductive transfer that improves generalization by using the domain

[1]The cross-entropy is a very common and popular cost function in machine learning (see [6] and references therein)

information contained in the training signals of related tasks as an inductive bias. 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 [24]. MTL has received a great deal of attention in the recent years [25]. 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 [26] or by modelling explicitly the links between tasks [27]. The other approach is through learning of common features for all desired tasks [28] 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 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 [29] for further details).

B. Our contribution

We first introduce an information-theoretic paradigm which provides the fundamental tradeoff between the *log-loss* (average risk) and the information rate of the features (statistical model complexity). 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 [11], [30], as described in Fig. 2, and prove its global convergence provided that some conditions hold. It worth to mention that our formulation, as a noisy source coding problem with side information at the decoder, provides an information-theoretic perspective to the MTL problem, which yields a valuable connection between the fields of machine learning and Shannon theory. In precise terms, the encoder aims at extracting relevant (common) information U from a data set X about labels $Y|Z = z$ needed for a collection of tasks $z \in \mathcal{Z}$ at the decoder. The function cost weights of each of these tasks will be defined to be the probability mass function of a randomly chosen index task variable 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 the model that compresses Y the best which is captured by Shannon mutual-information rate $I(U;X|Z = z)$. In the spirit of the Kolmogorov-Chaitin complexity [31] that is a measure of the regularities present in an object above and beyond pure randomness. 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. 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. All alphabets are assumed finite. $\mathbb{E}_{P_X}[\cdot]$ denotes the expectation and $|\mathcal{A}|$ indicates the cardinality of a set $\mathcal{A}$. $A \multimap B \multimap C$ 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 [32]: the *entropy* $H(X) \coloneqq \mathbb{E}_{P_X}[-\log P_X(X)]$, the *conditional entropy* $H(X|Y) \coloneqq \mathbb{E}_{P_{XY}}\left[-\log P_{X|Y}(X|Y)\right]$ and the *relative entropy*:

$$\mathcal{D}(P_X\|Q_X) = \begin{cases} \mathbb{E}_{P_X}\left[\log \dfrac{P_X(X)}{Q_X(X)}\right] & \text{if } P_X \ll Q_X \\ +\infty & \text{otherwise,} \end{cases} \tag{6}$$

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*: $I(X;Y) \coloneqq \mathcal{D}(P_{XY}\|P_XP_Y)$. 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

Let (X, Y, Z) be RVs with joint probability mass function P_{XYZ}. A soft-encoder $P_{U|X}$ wishes to extract from X information about a collection of labels $Y|Z = z$ with $z \in \mathcal{Z}$ while the randomly chosen task with index Z is available only at the decoder, as shown in Fig. 2.

Following our previous discussions right after (5), the optimal soft-decoder $P_{Y|UZ}$ will depend on the selected $(P_{U|X}, P_{XYZ})$ and is given by

$$P_{Y|UZ} = \frac{\sum_x P_{U|X}P_{XYZ}}{\sum_x P_{U|X}P_{XZ}}. \tag{7}$$

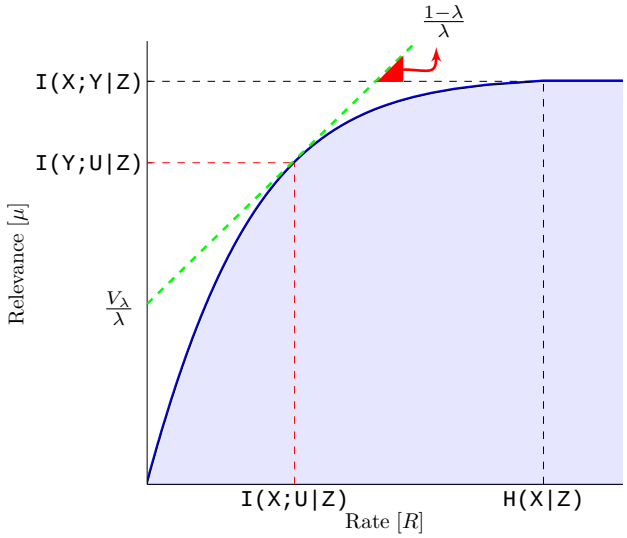

Fig. 3. A relevance-rate region with its supporting hyperplane (12).

We focus on the average *log-loss risk* that coincides with the conditional entropy:

$$\mathbb{E}_{P_{XYZ}P_{U|X}}[-\log P_{Y|UZ}(Y|UZ)] = H(Y|UZ). \quad (8)$$

Finding the encoder that minimizes the average log-loss is equivalent to search for the encoder maximizing the mutual (relevance) information $I(Y;U|Z) = H(Y|Z) - H(Y|UZ)$. As a consequence, we can focus on maximizing the relevance (mutual information) $I(Y;U|Z)$ subject to a given complexity (Shannon rate) $I(X;U|Z)$. As a matter of fact, it has been shown in [11, Theorem 1 with $\mu_2 = 0$, $L = 1$] that this tradeoff corresponds to the best possible asymptotically (over the block-length) achievable tradeoff between the multi-letter relevance and the compression rate.

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

$$\mathcal{R} \coloneqq \big\{(\mu, R) \in \mathbb{R}^2_{\geq 0} : \exists\, P_{U|X} \text{ s.t. } R \geq I(X;U|Z), \\ \mu \leq I(Y;U|Z), \quad U \multimap X \multimap (Y,Z)\big\}, \quad (9)$$

and the corresponding relevance-rate function is defined by

$$L(R, P_{XYZ}) = \max\big\{\mu : (R,\mu) \in \mathcal{R}\big\}, \quad (10)$$
$$= \max_{P_{U|X}:\, I(U;X|Z)\leq R} I(U;Y|Z). \quad (11)$$

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

Proof: See Appendix A. ∎

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 this problem is well-defined because we are attempting to maximize a continuous function over a compact set. An example of the relevance-rate region can be seen in Fig. 3. The relevance-rate function quantifies the maximum mutual information (conditioned on Z) between Y and the generated description U (using solely X) when a bound on the description complexity (rate) is imposed on U.

Although the optimization involved in (11) does not lead to a convex problem, the properties of $\mathcal{R}$ allows us to characterize the optimal tradeoff between compression and relevance rates using *supporting hyperplanes* [5]. As it is well known, any closed and convex set can be characterized from all its supporting hyperplanes [33]. A supporting hyperplane for $\mathcal{R}$ with parameter λ can be written as:

$$V_\lambda \coloneqq \max_{P_{U|X}} \lambda I(Y;U|Z) - (1-\lambda) I(X;U|Z). \quad (12)$$

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 (12) requires knowledge of the underlying distribution P_{XYZ}. 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: $\{(x_1,y_1,z_1),\ldots,(x_n,y_n,z_n)\}$ sampled according to the unknown distribution P_{XYZ}. In Section IV, we will study some supervised learning setups where expression (12) together with the iterative algorithm described below will serve as a supervised objective to guide multi-task learning.

C. An iterative optimization algorithm

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

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

Clearly, we can write

$$f(\lambda, P_{U|X}) = \sum_{z\in\mathcal{Z}} P_Z(z)\big[\lambda I(Y;U|Z=z) \\ - (1-\lambda) I(X;U|Z=z)\big], \quad (14)$$

where we see the effect of the weights $P_Z(z)$ associated with each task. *Data Processing Inequality* [34, sec. 2.3] allows to conclude that the only allowable values are $0 \leq \mu \leq I(X;Y|Z)$. 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 (12), and evaluating the corresponding mutual informations $I^{*,\lambda}(X;U|Z)$ and $I^{*,\lambda}(Y;U|Z)$ based on:

$$P^{*,\lambda}_{U|X} \coloneqq \arg\max_{P_{U|X}} f(\lambda, P_{U|X}). \quad (15)$$

By relying on the Markov chain $U \multimap (X,Z) \multimap Y$ implied from $U \multimap X \multimap (Y,Z)$, the function $f(\lambda, P_{U|X})$ writes as:

$$f(\lambda, P_{U|X}) = (2\lambda - 1) I(X;U|Z) - \lambda I(X;U|Y,Z). \quad (16)$$

Depending on the value of λ, it is appropriate to define the algorithm in two different ways. This is similar to the approach in [22]. If $\lambda \in [0, 0.5]$, both terms of (16) 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 relevant 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.

ALGORITHM 1: Information Bottleneck with side information.

Input: P_{XYZ}, $P_{U|X}^{(0)}$, $\lambda \in [0,1]$, $\epsilon > 0$.
Output: $P_{U|X}^{*,\lambda}$.

Initialize $n \coloneqq 0$, $I_\lambda^{(0)} \coloneqq +\infty$, $F_\lambda^{(0)} \coloneqq 0$.
while $I_\lambda^{(n)} - F_\lambda^{(n)} > \epsilon$ **do**

Compute

$$Q_{U|YZ}^{(n+1)} \coloneqq \sum_x P_{U|X}^{(n)} P_{X|YZ}, \quad Q_{X|ZU}^{(n+1)} \coloneqq \frac{P_{U|X}^{(n)} P_{X|Z}}{\sum_{x'} P_{U|X'}^{(n)} P_{X'|Z}},$$

$$P_{U|X}^{(n+1)} \coloneqq k_x \cdot \exp\left\{ \frac{2\lambda - 1}{\lambda} \sum_z P_{Z|X} \log(Q_{X|ZU}^{(n+1)}) + \sum_{y,z} P_{Y|XZ} P_{Z|X} \log(Q_{U|YZ}^{(n+1)}) \right\},$$

$$F_\lambda^{(n+1)} \coloneqq (2\lambda - 1) \sum_{x,z,u} P_{U|X}^{(n+1)} P_{XZ} \log\left(\frac{Q_{X|ZU}^{(n+1)}}{P_{X|Z}} \right) - \lambda \sum_{x,y,z,u} P_{U|X}^{(n+1)} P_{XYZ} \log\left(\frac{P_{U|X}^{(n+1)}}{Q_{U|YZ}^{(n+1)}} \right),$$

$$I_\lambda^{(n+1)} \coloneqq \max_u (2\lambda - 1) \sum_{x,z} P_{XZ} \log\left(\frac{Q_{X|ZU}^{(n+1)}}{P_{X|Z}} \right) - \lambda \sum_{x,y,z} P_{XYZ} \log\left(\frac{P_{U|X}^{(n)}}{Q_{U|YZ}^{(n+1)}} \right),$$

Update $n \coloneqq n + 1$.
end while
Report $P_{U|X}^{*,\lambda} = P_{U|X}^{(n)}$.

III. Algorithm Analysis

The problem of finding the global maximum is not convex because $P_{U|X} \mapsto f(\lambda, P_{U|X})$ is not concave. As a consequence, we cannot expect to have an efficient procedure that allow us to finding the global maximum of the problem. The algorithm proposed is a variant of the BA algorithm [16], [17] which is based on solid theoretical grounds and guarantee global optimum convergence results when the optimization problem is convex (e.g. the capacity and rate-distortion functions). Although our optimization problem is not convex and thus, the general convergence to the global optimum cannot be guaranteed, we derive results regarding the convergence to the global maximum provided that some additional conditions are fulfilled. Our results are inspired from seminal works in [20]–[22]. A convergence rate result is also derived in Appendix D.

A. Algorithm summary

We first study the algorithms expressions in further detail. Eq. (16) can be expanded as:

$$f(\lambda, P_{U|X}) = (2\lambda - 1) \sum_{x,z,u} P_{U|X} P_{XZ} \log\left(\frac{P_{X|ZU}}{P_{X|Z}} \right) - \lambda \sum_{x,y,z,u} P_{U|X} P_{XYZ} \log\left(\frac{P_{U|X}}{P_{U|YZ}} \right). \quad (17)$$

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

$$F(\lambda, P_{U|X}, Q_{U|YZ}, Q_{X|ZU}) \coloneqq (2\lambda - 1) \sum_{x,z,u} P_{U|X} P_{XZ} \log\left(\frac{Q_{X|ZU}}{P_{X|Z}} \right) - \lambda \sum_{x,y,z,u} P_{U|X} P_{XYZ} \log\left(\frac{P_{U|X}}{Q_{U|YZ}} \right), \quad (18)$$

where $Q_{U|YZ}, Q_{X|ZU}$ 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) $f(\lambda, P_{U|X}) \geq F(\lambda, P_{U|X}, Q_{U|YZ}, Q_{X|ZU})$, and equality is achieved iff $Q_{U|YZ} = P_{U|YZ}$ $\forall\ (y,z) \in \mathcal{Y} \times \mathcal{Z}$ and $Q_{X|ZU} = P_{X|ZU}$ $\forall (z,u) \in \mathcal{Z} \times \mathcal{U}$.
2) The value V_λ satisfies:
$$V_\lambda = \max_{P_{U|X}} \max_{Q_{U|YZ}, Q_{Y|ZU}} F(\lambda, P_{U|X}, Q_{U|YZ}, Q_{X|ZU}). \quad (19)$$
3) For any $Q_{U|YZ}, Q_{X|ZU}$ and $\lambda \in (0.5, 1]$, $P_{U|X} \mapsto F(\lambda, P_{U|X}, Q_{U|YZ}, Q_{X|ZU})$ is concave and achieves its maximum provided that:
$$P_{U|X} = k_x \cdot \exp\Big\{ \frac{2\lambda - 1}{\lambda} \sum_{z \in \mathcal{Z}} P_{Z|X} \log(Q_{X|ZU}) + \sum_{(y,z) \in \mathcal{Y} \times \mathcal{Z}} P_{Y|XZ} P_{Z|X} \log(Q_{U|YZ}) \Big\}, \quad (20)$$
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:
$$f(\lambda, P_{U|X}) - F(\lambda, P_{U|X}, Q_{U|YZ}, Q_{X|ZU}) = (2\lambda - 1) \sum_{x,y,z,u} P_{U|X} P_{XYZ} \log\left(\frac{P_{X|ZU}}{P_{X|Z}} \right) - \lambda \sum_{x,y,z,u} P_{U|X} P_{XYZ} \log\left(\frac{P_{U|X}}{P_{U|YZ}} \right) - (2\lambda - 1) \sum_{x,y,z,u} P_{U|X} P_{XYZ} \log\left(\frac{Q_{X|ZU}}{P_{X|Z}} \right)$$

$$+\lambda \sum_{x,y,z,u} P_{U|X} P_{XYZ} \log\left(\frac{P_{U|X}}{Q_{U|YZ}}\right) \quad (21)$$

$$= \lambda \sum_{y,z} P_{YZ}\, \mathcal{D}(P_{U|Y,Z} \| Q_{U|YZ}) + (2\lambda - 1) \sum_{z,u} P_{ZU}\, \mathcal{D}(P_{X|ZU} \| Q_{X|ZU}) \geq 0 \quad (22)$$

with equality iff $Q_{U|YZ} = P_{U|YZ}\ \forall\ (y,z) \in \mathcal{Y} \times \mathcal{Z}$ and $Q_{X|ZU} = P_{X|ZU}\ \forall\ (z,u) \in \mathcal{Z} \times \mathcal{U}$. This is easily seen from the properties of relative entropy [34, sec. 2.3].

2) The claim follows by combining the previous claim with (12).

3) Every pmf satisfies $\sum_u P_{U|X} = 1$. Then, using Lagrange multipliers $c_x,\ x \in \mathcal{X}$:

$$\frac{\partial[F + \sum_x c_x(\sum_u P_{U|X} - 1)]}{\partial P_{U|X}} = c_x - \lambda P_X \log(e P_{U|X}) + (2\lambda - 1) \sum_z P_{XZ} \log\left(Q_{X|ZU}\right) + \lambda \sum_{y,z} P_{XYZ} \log(Q_{U|YZ}) = 0 \quad (23)$$

from which we immediately recover (20). 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:

$$\frac{\partial^2 F(\lambda, P_{U|X}, Q_{U|YZ}, Q_{Y|ZU})}{\partial P_{U|X}^2} = -\frac{\lambda P_X \log(e)}{P_{U|X}} \leq 0. \quad (24)$$

■

We observe that the function $F(\lambda, P_{U|X}, Q_{U|YZ}, Q_{X|ZU})$ 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, $P_{U|X} \mapsto F(\lambda, P_{U|X}, Q_{U|YZ}, Q_{X|ZU})$ is concave for each $(Q_{U|YZ}, Q_{X|ZU})$, 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_λ. This is the case in Algorithm 1, where we perform an iterative maximization process on both arguments: $P_{U|X}$ and $(Q_{U|YZ}, Q_{X|ZU})$. For a given $\lambda \in (0.5, 1]$, starting from an initial condition $P_{U|X}^{(0)}$, and according to 2) in Lemma 2, we search for $Q_{U|YZ}^{(1)}, Q_{X|ZU}^{(1)}$ such that the maximum of $F(\lambda, P_{U|X}^{(0)}, Q_{U|YZ}, Q_{X|ZU})$ is achieved, for fixed $P_{U|X}^{(0)}$. Next, from 3) in the previous lemma, we find $P_{U|X}^{(1)}$ as the argument that maximizes $F(\lambda, P_{U|X}, Q_{U|YZ}^{(1)}, Q_{X|ZU}^{(1)})$. 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 $F(\lambda, P_{U|X}^{(n)}, Q_{U|YZ}^{(n)}, Q_{X|ZU}^{(n)})$ is monotone non-decreasing. This clearly guarantees the convergence. In the sequel, we further study this process in detail.

B. Convergence analysis

For sake of simplicity, let us assume that the optimal point $P_{U|X}^{*,\lambda}$ is unique. Define $F_\lambda^{(n)} \coloneqq F(\lambda, P_{U|X}^{(n)}, Q_{U|YZ}^{(n)}, Q_{X|ZU}^{(n)})$. From the previous section we know that $F_\lambda^{(n+1)} \geq F_\lambda^{(n)}$. Moreover, from 2) in Lemma 2, $V_\lambda \geq F_\lambda^{(n)}$ for all n. However, there is no guarantee that $V_\lambda = F_\lambda^{(\infty)}$. 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 δ-superlevel set (see [20] for further details).

Definition 2: The δ-superlevel set is defined as the set:

$$G_{\delta,\lambda} \coloneqq \left\{ P_{U|X} : \mathcal{X} \to \mathcal{P}(\mathcal{U}) \middle|\ f(\lambda, P_{U|X}) \geq \delta \right\}. \quad (25)$$

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 $P_{U|X}^{(n+1)}$ lies in $H_{\delta,\lambda}(P_{U|X}^{(n)})$ for all k such that $P_{U|X}^{(n)} \in G_{\delta,\lambda}$.

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

$$F_\lambda^{(n+1)} \geq F(\lambda, P_{U|X}^{(n)}, Q_{U|YZ}^{(n+1)}, Q_{X|ZU}^{(n+1)}) = f(\lambda, P_{U|X}^{(n)}). \quad (26)$$

For fixed $(Q_{U|YZ}^{(n+1)}, Q_{X|ZU}^{(n+1)})$ pmfs, we know that F is concave in argument $P_{U|X}$. Thus, $\tilde{G}_{\delta,\lambda}^n$ 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 $\tilde{G}_{\delta,\lambda}^n \subseteq H_{\delta,\lambda}(P_{U|X}^{(n)})$ and we conclude that $P_{U|X}^{(n+1)} \in H_{\delta,\lambda}(P_{U|X}^{(n)})$. ■

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

Lemma 4: Consider $\lambda \in (0.5, 1]$ and $P_{U|X}^{(0)} \in G_{\delta,\lambda}$ for a given value[2] of δ. If the optimal solution $P_{U|X}^{*,\lambda}$ lies in $H_{\delta,\lambda}(P_{U|X}^{(0)})$, the function $f(\lambda, P_{U|X})$ is concave in $H_{\delta,\lambda}(P_{U|X}^{(0)})$, then the following inequalities hold for every n:

$$V_\lambda \leq \lambda \sum_{x,y,z,u} P_{U|X}^{*,\lambda} P_{XYZ} \log\left(\frac{Q_{U|YZ}^{(n+1)}}{P_{U|X}^{(n)}}\right) + (2\lambda - 1) \sum_{x,z,u} P_{U|X}^{*,\lambda} P_{XZ} \log\left(\frac{Q_{X|ZU}^{(n+1)}}{P_{X|Z}}\right), \quad (27)$$

$$V_\lambda - F_\lambda^{(n+1)} \leq \lambda \sum_{x,u} P_{U|X}^{*,\lambda} P_X \log\left(\frac{P_{U|X}^{(n+1)}}{P_{U|X}^{(n)}}\right). \quad (28)$$

Proof: See Appendix B. ■

Theorem 1: Consider $\lambda \in (0.5, 1]$ and $P_{U|X}^{(0)} \in G_{\delta,\lambda}$ for a given value of δ. If the optimal solution $P_{U|X}^{*,\lambda}$ lies

[2]It is easy to show that we can always find a value of δ such that this condition is satisfied.

in $H_{\delta,\lambda}(P_{U|X}^{(0)})$ and the function $f(\lambda, P_{U|X})$ is concave in $H_{\delta,\lambda}(P_{U|X}^{(0)})$ and $P_{U|X}^{(0)}$ is such that $|\text{supp}(P_{U|X}^{(0)})| = |\mathcal{U}|$, then:

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

Proof:

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

$$\sum_{n=0}^{N-1} V_\lambda - F_\lambda^{(n+1)} \leq \lambda \sum_{x,u} P_{U|X}^{*,\lambda} P_X \log\left(\frac{P_{U|X}^{(N)}}{P_{U|X}^{(0)}}\right) \tag{29}$$

$$= \lambda\left(\mathbb{E}_X\left[\mathcal{D}(P_{U|X}^{*,\lambda}\|P_{U|X}^{(0)})\right] - \mathbb{E}_X\left[\mathcal{D}(P_{U|X}^{*,\lambda}\|P_{U|X}^{(N)})\right]\right) \tag{30}$$

$$\leq \lambda \mathbb{E}_X\left[\mathcal{D}(P_{U|X}^{*,\lambda}\|P_{U|X}^{(0)})\right], \tag{31}$$

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

■

As $f(\lambda, P_{U|X})$ is not globally concave, the δ-superlevel set $G_{\delta,\lambda}$ neither convex but may not also be connected. By Lemma 3, the algorithm proposed stays in the path-connected component $H_{\delta,\lambda}(P_{U|X}^{(0)})$ which is determined by the initial condition. If the the optimal point $P_{U|X}^{*,\lambda}$ is contained in the right path-connected component $H_{\delta,\lambda}(P_{U|X}^{(0)})$ and $f(\lambda, P_{U|X})$ is locally concave around the optimal point $P_{U|X}^{*,\lambda}$, which is something reasonable to expect because of the smoothness of $f(\lambda, P_{U|X})$. The above results give positive answers regarding the convergence of the algorithm to the optimal point $P_{U|X}^{*,\lambda}$. 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 initial conditions and keep the one that provides the largest value of F after stopping condition is met.

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 $P_{U|X}^{*,\lambda}$, we have

$$\begin{aligned} &\alpha^*(\lambda, u) = V_\lambda, \quad \text{for } u \in \mathcal{U} \text{ such that } P_{U|X}^{*,\lambda} > 0 \text{ and } \forall x \in \mathcal{X} \\ &\alpha^*(\lambda, u) \leq V_\lambda, \quad \text{otherwise,} \end{aligned} \tag{32}$$

where

$$\alpha^*(\lambda, u) \coloneqq (2\lambda - 1) \sum_{(x,z)\in\mathcal{X}\times\mathcal{Z}} P_{XZ} \log\left(\frac{Q_{X|ZU}^{*,\lambda}}{P_{X|Z}}\right) + \lambda \sum_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} P_{XYZ} \log\left(\frac{Q_{U|YZ}^{*,\lambda}}{P_{U|X}^{*,\lambda}}\right). \tag{33}$$

Proof: See Appendix C. ■

It is interesting to observe that the optimal solution $P_{U|X}^{*,\lambda}$ is such that for each value of $u \in \mathcal{U}$ where $P_{U|X}^{*,\lambda} > 0$ the value of $\alpha^*(\lambda, u)$ is constant and equal to the maximum value V_λ. 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 [35]. In particular, we have:

$$V_\lambda = \max_{u\in\mathcal{U}} \alpha^*(\lambda, u). \tag{34}$$

From these results, we can consider the quantity:

$$I_\lambda^{(n+1)} \coloneqq \max_{u\in\mathcal{U}} (2\lambda - 1) \sum_{(x,z)\in\mathcal{X}\times\mathcal{Z}} P_{XZ} \log\left(\frac{Q_{X|ZU}^{(n+1)}}{P_{X|Z}}\right) + \lambda \sum_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} P_{XYZ} \log\left(\frac{Q_{U|YZ}^{(n+1)}}{P_{U|X}^{(n)}}\right). \tag{35}$$

It is clear from (27) 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 $P_{U|X}^{*,\lambda}$ 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 (11)) 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 [13]. In this problem, the pmf P_{XYZ} is a probability measure corresponding to a source (X, Y, Z) such that $Y \multimap X \multimap Z$ 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 [13] that the relevance-rate region is given by the convex hull [33] of the following region:

$$\mathcal{R}_b \coloneqq \big\{(R, \mu) : R \geq I(X;U|Z),\ \mu \leq I(Y;U|Z), \text{ with } P_{U|X} = \text{BSC}(r)\ \ \forall r \in [0, 0.5]\big\}, \tag{36}$$

where $\text{BSC}(r)$ 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

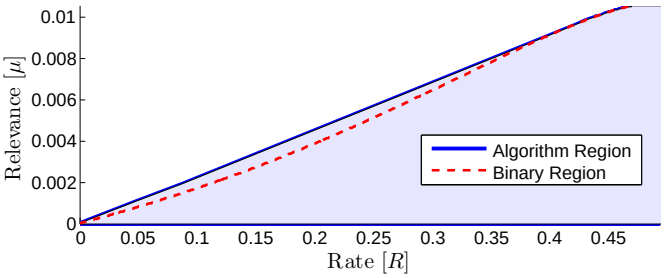

Fig. 4. Rate-relevance region corresponding to a double symmetric binary source.

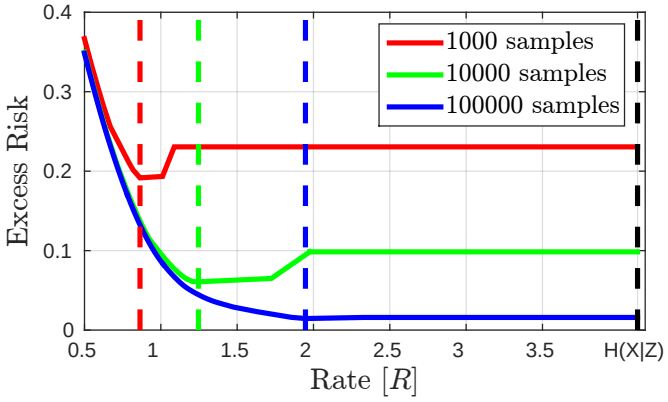

Fig. 5. Excess risk (39) as a function of the information rate.

pairs in $\mathcal{R}$ for an arbitrary pmf P_{XYZ}. 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_{XYZ}. In order to test the suitability of the algorithm for this task we used it 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_{XYZ} is not known but it is estimated from training samples. Indeed, Shamir *et al.* [36] have already showed evidence that this term can help to prevent "overfitting" and this idea was also exploited in [12], [37] 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. In this section, we provide numerical evidence that the desired regularization effects hold in our multi-task learning setup.

Consider a multi-task supervised classification problem and define the average cross-entropy risk:

$$\text{Risk}\big(P_{U|X}, P_{\hat{Y}|UZ}\big) := \mathbb{E}_{P_{XYZ}P_{U|X}}[-\log P_{\hat{Y}|UZ}(Y|UZ)] \tag{37}$$

with respect to $P_{U|X}$ and $P_{\hat{Y}|UZ}$. We notice that this risk is not necessarily equivalent to the classification error. However, it is easy to check that it is an appropriate surrogate:

$$\mathbb{P}\big(Y \neq \hat{Y}\big) \leq 1 - 2^{-\text{Risk}\left(P_{U|X}, P_{\hat{Y}|UZ}\right)}. \tag{38}$$

Finding the optimal encoder in (37) requires knowledge of the underlying distribution P_{XYZ}. From a practical perspective, as the input to the proposed algorithm, we will use the data sampling distribution $\hat{P}_{XYZ}$ based on n training labeled examples. By introducing the rate constraint (or penalization), the optimization problem is reduced to optimizing $L(R, \hat{P}_{XYZ})$ in (11) from which the resulting encoder $\hat{P}^{*,\lambda}_{U|X}$ is derived while the decoder $\hat{P}^{*,\lambda}_{\hat{Y}|UZ}$ follows from expression[3] (7). The measure of merit will be the *Excess-risk*:

$$\text{Excess-risk} := \text{Risk}\big(\hat{P}^{*,\lambda}_{U|X}, \hat{P}^{*,\lambda}_{\hat{Y}|UZ}\big) - H(Y|XZ) \tag{39}$$

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

Experiments will be performed by using synthetic data with alphabets $|\mathcal{X}| = 64$, $|\mathcal{Y}| = 4$, $|\mathcal{Z}| = 2$. The random variable Z is assumed to follow a *Bernoulli* distribution with random parameter $p \in [0, 1]$ while the joint distribution $P_{XY|Z=z}$ is defined as a *Restrict Boltzmann Machine* (see [6] for further details) with parameters randomly drawn for each $z \in \mathcal{Z}$.

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_{lim}, such that for each $R \geq R_{\text{lim}}$, the excess-risk remains constant with value $\hat{I}(X;Y|Z)$. 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_{opt} which provides the lowest excess-risk in (39). 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.

C. Hierarchical text categorization

The high dimensionality of texts can become a severe deterrent in applying complex learners like support vector

[3]Notice that when the decoder is chosen as in (7), then $\text{Risk}\big(P_{U|X}, P_{\hat{Y}|UZ}\big) = H(Y|UZ) \geq H(Y|XZ)$, where the inequality is a consequence of the Markov chain $U \multimap X \multimap (Y, Z)$.

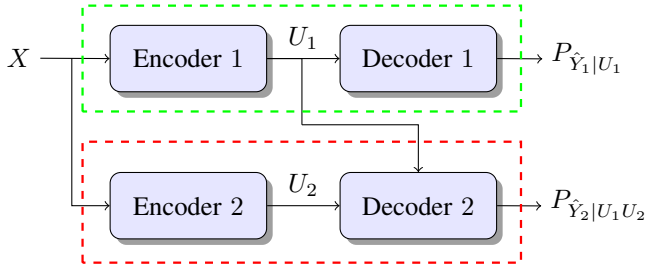

Fig. 6. Hierarchical text categorization and multi-task learning.

machines (SVM) [6] to the task of text classification. Word clustering is a powerful alternative to feature selection for reducing the dimensionality of text [8], [38]. 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 [19], [39]. 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_{XY_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 $P_{XY_1U_1}$. 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 [40]. Imposing the Markov chain $U_1 \multimap X \multimap Y_2$, the resulting joint pmfs are given by $P_{XY_2U_1} = P_{U_1|X}P_{X|Y_2}P_{Y_2}$ and

$$P_{XY_1} = \sum_{y_2 \in \mathcal{Y}_2} P_{Y_1|Y_2} P_{X|Y_2} P_{Y_2}. \tag{40}$$

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 [38] for details):

$$\hat{y}_2(d) = \arg\max_{y \in \mathcal{Y}_2} P_{Y_2} \prod_{u_1,u_2} \left(P_{U_1U_2|Y_2}\right)^{n(u_1,u_2,d)} \tag{41}$$

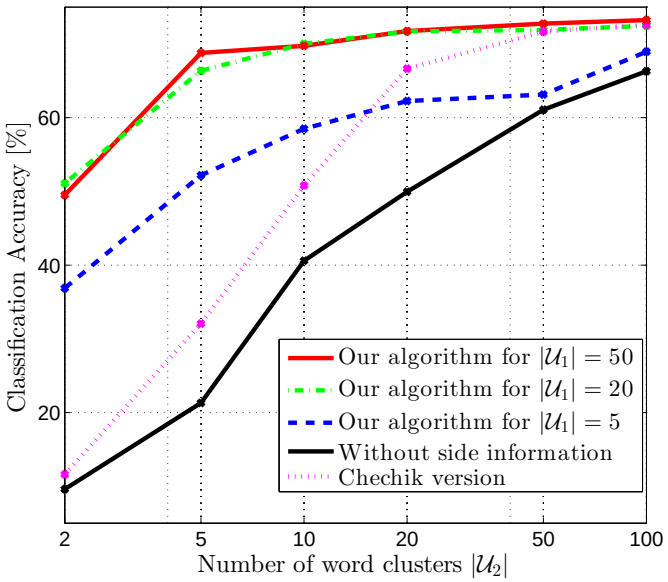

Fig. 7. Classification Accuracy in the hierarchical text categorization problem.

$$\equiv \arg\max_{y \in \mathcal{Y}_2} \log\left(P_{Y_2}\right) + \sum_{u_1,u_2} n(u_1,u_2,d) \log P_{U_1U_2|Y_2}, \tag{42}$$

where

$$P_{U_1U_2|Y_2} = \frac{\sum_x P_{XY_2U_1} P_{U_2|X}}{P_{Y_2}}, \tag{43}$$

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:

$$u_1(x) := \arg\max_u P_{U_1|X}(u|x),\ u_2(x) := \arg\max_u P_{U_2|X}(u|x). \tag{44}$$

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 [19]. It is interesting to mention that the single-task setting and the one in [19] can be covered using the proposed algorithm. In particular, for the single-task setup, we estimate the final class with P_{XY_2} as the input of the proposed algorithm (Z is a degenerate RV). Our setting and the one in [19] 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.

Remark 1: There exists a strong relationship between our objective and the one in [19] 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:

$$f(\lambda, P_{U|X}) = \lambda I(Y;U|g(Y)) - \bar{\lambda} I(X;U|g(Y)) \tag{45}$$
$$= \lambda \left[I(Y;U) - I(g(Y);U)\right] - \bar{\lambda}\left[I(X;U) - I(g(Y);U)\right] \tag{46}$$
$$= \lambda I(Y;U) - (2\lambda - 1) I(g(Y);U) - \bar{\lambda} I(X;U), \tag{47}$$

where $\bar{\lambda} = 1 - \lambda$ and $f(\lambda, P_{U|X})$ depends on the source via the marginal distributions: P_{XY} and P_{XZ}. This expression is the cost function proposed in [19] 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 Multi-task learning in which an encoder builds a common representation intended to several related tasks. We derived an iterative learning algorithm from the principle of compression-based regularization that uses compression 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. An application to hierarchical text categorization was also considered.

At present, 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 multi-task learning setups, besides the hierarchical text categorization one, should also deserve additional efforts.

Appendix

Appendix A: Proof of Lemma 1

We show that the region $\mathcal{R}$ is closed, convex and we bound the cardinality of the RV U. A region $\mathcal{R}$ is closed iff it contains the limit of every converging sequence whose terms lie in $\mathcal{R}$. Let $(\mu^{(k)}, R^{(k)}) \in \mathcal{R}$ such that $(\mu^{(k)}, R^{(k)}) \overset{k\to\infty}{\longrightarrow} (\mu, R)$. As $(\mu^{(k)}, R^{(k)}) \in \mathcal{R}$ we have that exists $P^{(k)}_{U|X}$ such that:

$$R^{(k)} \geq I^{(k)}(X;U|Z), \;\; \mu^{(k)} \leq I^{(k)}(Y;U|Z). \tag{48}$$

We have then a sequence of conditional probability distributions $\left\{P^{(k)}_{U|X}\right\}_{k=1}^{\infty}$. As this sequence is in a compact set (i.e. the set of all conditional PDs with alphabets $\mathcal{X}$ and $\mathcal{U}$) it should exists a converging subsequence $P^{(k_n)}_{U|X}$ with $n \in \mathbb{N}$ and limiting point $P^{\lim}_{U|X}$. Consider the subsequence $(\mu^{(k_n)}, R^{(k_n)})$, it is straightforward to check that for any $\epsilon > 0$ and n large enough, $|R - R^{(k_n)}| < \epsilon$ and $|\mu - \mu^{(k_n)}| < \epsilon$. Then, we have:

$$R \geq I^{\lim}(X;U|Z) - \epsilon, \;\; \mu \leq I^{\lim}(Y;U|Z) + \epsilon. \tag{49}$$

As ϵ is arbitrary we conclude that $(\mu, R) \in \mathcal{R}$.

For the convexity analysis consider $(R_1, \mu_1), (R_2, \mu_2) \in \mathcal{R}$. We will prove $\theta(R_1, \mu_1) + (1-\theta)(R_2, \mu_2) \in \mathcal{R} \;\forall\; \theta \in [0, 1]$.

If $(R_1, \mu_1) \in \mathcal{R}$, $R_1 \geq I(X;V_1|Z)$ and $\mu_1 \leq I(Y;V_1|Z)$ for a conditional distribution $P_{V_1|X}$. Similarly, if $(R_2, \mu_2) \in \mathcal{R}$, $R_2 \geq I(X;V_2|Z)$ and $\mu_2 \leq I(Y;V_2|Z)$ with distribution $P_{V_2|X}$. We define $T \sim \text{Ber}(\theta)$ (Bernoulli RV with parameter θ) independent of (X, Y, Z, V_1, V_2), and the RV:

$$V = \begin{cases} V_1, & \text{If } \; T = 1 \\ V_2, & \text{If } \; T = 0. \end{cases} \tag{50}$$

Then,

$$\theta R_1 + (1-\theta) R_2$$
$$\geq \theta I(X;V_1|Z) + (1-\theta) I(X;V_2|Z) \tag{51}$$
$$= \theta I(X;V|Z,T=1) + (1-\theta) I(X;V|Z,T=0) \tag{52}$$
$$= I(X;V,T|Z) \tag{53}$$
$$= I(X;U|Z), \tag{54}$$

where $U = VT$. The same happens with the relevance:

$$\theta \mu_1 + (1-\theta)\mu_2$$
$$\leq \theta I(Y;V_1|Z) + (1-\theta) I(Y;V_2|Z) \tag{55}$$
$$= \theta I(Y;V|Z,T=1) + (1-\theta) I(Y;V|Z,T=0) \tag{56}$$
$$= I(Y;V,T|Z) \tag{57}$$
$$= I(Y;U|Z). \tag{58}$$

It is necessary to show that U have the Markov property $U \multimap X \multimap (Y, Z)$:

$$I(Y,Z;V,T|X) = I(Y,Z;V|X,T) \tag{59}$$
$$= \theta I(Z;V|X,T=1) + (1-\theta) I(Z;V|X,T=0) \tag{60}$$
$$= \theta I(Z;V_1|X) + (1-\theta) I(Z;V_2|X) = 0. \tag{61}$$

It is clear then that $\mathcal{R}$ is convex. Now, we will show that the cardinality of RV U can be bounded $|\mathcal{U}| \leq |\mathcal{X}| + 1$ without loss of generality. This follows easily from the Support Lemma [34, app. C], witch follows from Carathéodry Theorem.

Lemma 6 (Support Lemma): Let d functions $g_1, g_2, \ldots, g_d$ of conditional probabilities $P_{X|U}$. Then, for all U exists U' with cardinality $|\mathcal{U}'| \leq d$ s.t. $\mathbb{E}_U[g(P_{X|U})] = \mathbb{E}_{U'}[g(P_{X|U'})]$.

For our problem the choice of these functions is done in order to preserve *Markov Chains* and the mutual information expressions for the bounds on the rate and the relevance. It is an almost trivial exercise to check that the functions amount to a quantity of $|\mathcal{X}| + 1$ atoms.

Appendix B: Proof of Lemma 4

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

Lemma 7: 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

$$\mathcal{L}[P_a, P_b] = \sum_{(x,y,z,u)} P_a P_{X,Y,Z} \big[\lambda \mathcal{D}\left(P_a \| P_b\right)$$
$$- \lambda \mathcal{D}\left(Q_{U|Y,Z}[P_a] \| Q_{U|YZ}[P_b]\right)$$
$$- (2\lambda - 1)\mathcal{D}\left(Q_{X|UZ}[P_a] \| Q_{X|U,Z}[P_b]\right)\big], \tag{62}$$

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

$$Q_{U|YZ}[P_i]=\sum_{x\in\mathcal{X}}P_iP_{X|YZ},\;\; Q_{X|ZU}[P_i]=\frac{P_iP_{X|Z}}{\sum_x P_iP_{X|Z}}. \tag{63}$$

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: $[f(x)\log(f(x))]'=f'(x)\log\left(ef(x)\right)$ and obtain:

$$\begin{aligned}\frac{\partial f(\lambda,P_{U|X})}{\partial P_{U|X}}&=(2\lambda-1)P_X\log\left(eP_{U|X}\right)\\&-(2\lambda-1)\sum_z P_{X,Z}\log\left(eP_{U|Z}\right)\\&-\lambda\,P_X\log\left(eP_{U|X}\right)+\lambda\sum_{y,z}P_{XYZ}\log\left(eP_{U|YZ}\right) \quad (64)\\&=(2\lambda-1)\sum_z P_{XZ}\log\left(\frac{P_{U|X}}{P_{U|Z}}\right)\\&-\lambda\sum_{y,z}P_{XYZ}\log\left(\frac{P_{U|X}}{P_{U|YZ}}\right) \quad (65)\\&=(2\lambda-1)\sum_z P_{XZ}\log\left(\frac{P_{X|ZU}}{P_{X|Z}}\right)\\&-\lambda\sum_{y,z}P_{XYZ}\log\left(\frac{P_{U|X}}{P_{U|YZ}}\right). \quad (66)\end{aligned}$$

Note that

$$f(\lambda,P_{U|X})=\sum_{x,u}P_{U|X}\frac{\partial f(\lambda,P_{U|X})}{\partial P_{U|X}}. \tag{67}$$

Then,

$$\sum_{x,u}P_b\left.\frac{\partial f(\lambda,P_{U|X})}{\partial P_{U|X}}\right|_{P_b}=f(\lambda,P_b). \tag{68}$$

Now, let us consider:

$$\begin{aligned}&\sum_{x,u}P_a\left.\frac{\partial f(\lambda,P_{U|X})}{\partial P_{U|X}}\right|_{P_b}\\&=f(\lambda,P_a)-(2\lambda-1)\sum_{x,z,u}P_aP_{XZ}\log\left(\frac{Q_{X|ZU}[P_a]}{Q_{X|ZU}[P_b]}\right)\\&+\lambda\sum_{x,u}P_aP_X\log\left(\frac{P_a}{P_b}\right)\\&-\lambda\sum_{x,y,z,u}P_aP_{XYZ}\log\left(\frac{Q_{U|YZ}[P_a]}{Q_{U|YZ}[P_b]}\right) \quad (69)\\&=f(\lambda,P_a)\\&+\sum_{x,y,z,u}P_aP_{XYZ}\left[-(2\lambda-1)\mathcal{D}\left(Q_{X|ZU}[P_a]\|Q_{X|ZU}[P_b]\right)\right.\\&\left.+\lambda\mathcal{D}\left(P_a\|P_b\right)-\lambda\mathcal{D}\left(Q_{U|YZ}[P_a]\|Q_{U|YZ}[P_b]\right)\right] \quad (70)\\&=f(\lambda,P_a)+\mathcal{L}[P_a,P_b]. \quad (71)\end{aligned}$$

Then,

$$\begin{aligned}\mathcal{L}[P_a,P_b]&=\sum_{x,u}P_a\left.\frac{\partial f(\lambda,P_{U|X})}{\partial P_{U|X}}\right|_{P_b}-f(\lambda,P_a) \quad (72)\\&=\sum_{x,u}(P_a-P_b)\left.\frac{\partial f(\lambda,P_{U|X})}{\partial P_{U|X}}\right|_{P_b}-f(\lambda,P_a)+f(\lambda,P_b). \quad (73)\end{aligned}$$

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

$$f(\lambda,P_a)\leq f(\lambda,P_b)+\sum_{x,u}\left.\frac{\partial f(\lambda,P_{U|X})}{\partial P_{U|X}}\right|_{P_b}(P_a-P_b), \tag{74}$$

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

Now we can proceed to the proof of Lemma 4. In order to show (27), we define the quantity $\mathcal{B}$:

$$\begin{aligned}\mathcal{B}\coloneqq(2\lambda-1)\sum_{x,z,u}P^{*,\lambda}_{U|X}P_{XZ}\log\left(\frac{Q^{(n+1)}_{X|ZU}}{P_{X|Z}}\right)\\+\lambda\sum_{x,y,z,u}P^{*,\lambda}_{U|X}P_{XYZ}\log\left(\frac{Q^{(n+1)}_{U|YZ}}{P^{(n)}_{U|X}}\right). \quad (75)\end{aligned}$$

Then, we can write:

$$\begin{aligned}V_\lambda&=(2\lambda-1)\sum_{x,z,u}P^{*,\lambda}_{U|X}P_{XZ}\log\left(\frac{Q^{*}_{X|ZU}}{P_{X|Z}}\right)\\&+\lambda\sum_{x,y,z,u}P^{*,\lambda}_{U|X}P_{XYZ}\log\left(\frac{Q^{*}_{U|YZ}}{P^{*,\lambda}_{U|X}}\right) \quad (76)\\&=\mathcal{B}+(2\lambda-1)\sum_{x,z,u}P^{*,\lambda}_{U|X}P_{XZ}\mathcal{D}\left(Q^{*}_{X|ZU}\|Q^{(n+1)}_{X|ZU}\right)\\&+\lambda\sum_{y,z}P_{YZ}\mathcal{D}\left(Q^{*}_{U|YZ}\|Q^{(n+1)}_{U|YZ}\right)\\&-\lambda\sum_{x}P_X\mathcal{D}\left(P^{*,\lambda}_{U|X}\|P^{(n)}_{U|X}\right) \quad (77)\\&=\mathcal{B}-\mathcal{L}[P^{*,\lambda}_{U|X},P^{(n)}_{U|X}]. \quad (78)\end{aligned}$$

Consider an integer $n\geq 1$ 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 7 to $\mathcal{L}[P^{*,\lambda}_{U|X},P^{(n)}_{U|X}]$ and conclude that $V_\lambda\leq\mathcal{B}$. We also define:

$$\begin{aligned}\gamma^{(n+1)}_{U|X}=\exp\left\{\frac{2\lambda-1}{\lambda}\sum_z P_{Z|X}\log(Q^{(n+1)}_{X|ZU})\right.\\\left.+\sum_{y,z}P_{Y|XZ}P_{Z|X}\log(Q^{(n+1)}_{U|YZ})\right\}, \quad (79)\end{aligned}$$

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

$$\mathcal{B}=\lambda\sum_{x,u}P^{*,\lambda}_{U|X}P_X\log\left(\frac{\gamma^{(n+1)}_{U|X}}{P^{(n)}_{U|X}}\right)+(2\lambda-1)H(X|Z). \tag{80}$$

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

$$F^{(n+1)}_\lambda=(2\lambda-1)\sum_{x,z,u}P^{(n+1)}_{U|X}P_{XZ}\log\left(\frac{Q^{(n+1)}_{X|ZU}}{P_{X|Z}}\right)$$

$$- \lambda \sum_{x,y,z,u} P_{U|X}^{(n+1)} P_{XYZ} \log \left(\frac{\gamma_{U|X}^{(n+1)}}{Q_{U|YZ}^{(n+1)} \sum_{u'} \gamma_{U'|X}^{(n+1)}} \right) \quad (81)$$

$$= (2\lambda - 1) \sum_{x,z,u} P_{U|X}^{(n+1)} P_{XZ} \log \left(\frac{Q_{X|ZU}^{(n+1)}}{P_{X|Z}} \right)$$

$$+ \lambda \sum_{x,y,z,u} P_{U|X}^{(n+1)} P_{XYZ} \log \left(Q_{U|YZ}^{(n+1)} \sum_{u'} \gamma_{U'|X}^{(n+1)} \right)$$

$$- (2\lambda - 1) \sum_{x,z,u} P_{U|X}^{(n+1)} P_{XZ} \log(Q_{X|ZU}^{(n+1)})$$

$$- \lambda \sum_{x,y,z,u} P_{U|X}^{(n+1)} P_{XYZ} \log(Q_{U|YZ}^{(n+1)}) \quad (82)$$

$$= \lambda \sum_{x} P_X \log \left(\sum_{u'} \gamma_{U'|X}^{(n+1)} \right) + (2\lambda - 1) H(X|Z). \quad (83)$$

Finally,

$$V_\lambda - F_\lambda^{(n+1)} \leq \mathcal{B} - \lambda \sum_{x} P_X \log \left(\sum_{u'} \gamma_{U'|X}^{(n+1)} \right) - (2\lambda - 1) H(X|Z) \quad (84)$$

$$= \lambda \sum_{x,u} P_{U|X}^{*,\lambda} P_X \log \left(\frac{P_{U|X}^{(n+1)}}{P_{U|X}^{(n)}} \right). \quad (85)$$

Appendix C: Proof of Lemma 5

The proofs is along the lines of Karush-Kuhn-Tucker (KKT) conditions [5]. 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 $(u, x) \in \mathcal{U} \times \mathcal{X}$:

$$\kappa_x \geq \frac{\partial f(\lambda, P_{U|X}^{*,\lambda})}{\partial P_{U|X}}. \quad (86)$$

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 (66) we have:

$$\frac{\partial f(\lambda, P_{U|X}^{*,\lambda})}{\partial P_{U|X}} = (2\lambda - 1) \sum_{z} P_{XZ} \log \left(\frac{Q_{X|ZU}^{*}}{P_{X|Z}} \right) - \lambda \sum_{y,z} P_{XYZ} \log \left(\frac{P_{U|X}^{*,\lambda}}{Q_{U|YZ}^{*}} \right). \quad (87)$$

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

$$\sum_{x} \kappa_x \geq (2\lambda - 1) \sum_{x,z} P_{XZ} \log \left(\frac{Q_{X|ZU}^{*}}{P_{X|Z}} \right) - \lambda \sum_{x,y,z} P_{XYZ} \log \left(\frac{P_{U|X}^{*,\lambda}}{Q_{U|YZ}^{*}} \right) \quad (88)$$

$$= \alpha^*(\lambda, u). \quad (89)$$

In a similar manner, from conditions 1) and 5) and using Eq. (67), we can write:

$$V_\lambda = f(\lambda, P_{U|X}^{*,\lambda}) = \sum_{x,u} P_{U|X}^{*,\lambda} \frac{\partial f(\lambda, P_{U|X}^{*,\lambda})}{\partial P_{U|X}} \quad (90)$$

$$= \sum_{x,u} P_{U|X}^{*,\lambda} (\kappa_x - \phi_{x,u}) \quad (91)$$

$$= \sum_{x} \kappa_x. \quad (92)$$

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. $P_{U|X}^{*,\lambda} > 0$.

Appendix D: Convergence rate

The speed of convergence can be obtained easily from the proof of Theorem 1.

Lemma 8: Consider $\lambda \in (0.5, 1]$ and $P_{U|X}^{(0)} \in G_{\delta,\lambda}$ for a given value of δ. If the optimal solution $P_{U|X}^{*,\lambda}$ lies in $H_{\delta,\lambda}(P_{U|X}^{(0)})$ and the function $f(\lambda, P_{U|X})$ is concave in $H_{\delta,\lambda}(P_{U|X}^{(0)})$ and $P_{U|X}^{(0)}$ is such that $|\text{supp}(P_{U|X}^{(0)})| = |\mathcal{U}|$, we have:

$$V_\lambda - F_\lambda^{(N)} \leq \frac{\lambda}{N} \cdot \mathbb{E}_X[\mathcal{D}(P_{U|X}^{*,\lambda} \| P_{U|X}^{(0)})]. \quad (93)$$

Proof: $F_\lambda^{(n)}$ is monotonically non-decreasing and is bounded by V_λ. Then,

$$\sum_{n=0}^{N-1} V_\lambda - F_\lambda^{(n+1)} \geq N(V_\lambda - F_\lambda^{(N)}). \quad (94)$$

The LHS term can be bounded using (31):

$$\lambda \mathbb{E}_X \left[\mathcal{D}(P_{U|X}^{*,\lambda} \| P_{U|X}^{(0)}) \right] \geq N(V_\lambda - F_\lambda^{(N)}). \quad (95)$$

Finally,

$$V_\lambda - F_\lambda^{(N)} \leq \frac{\lambda}{N} \cdot \mathbb{E}_X[\mathcal{D}(P_{U|X}^{*,\lambda} \| P_{U|X}^{(0)})]. \quad (96)$$

■

We can see that the approximation error $V_\lambda - F_\lambda^{(N)}$ is inversely proportional to the number of iterations and the algorithm has a rate of convergence of at least order $1/N$.

References

[1] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," in *Claude Elwood Shannon: collected papers*, N. J. A. Sloane and A. D. Wyner, Eds. IEEE Press, 1993, pp. 325–350.

[2] 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.

[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] 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.

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

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

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

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

[14] 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

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

[16] 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.

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

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

[19] 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.

[20] 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.

[21] 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.

[22] 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.

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

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

[25] 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

[26] 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.

[27] 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

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

[29] 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.

[30] 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.

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

[32] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

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

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

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

[36] 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

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

[38] 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.

[39] 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.

[40] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

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