Information Bottleneck and Representation Learning

Pablo Piantanida, Leonardo Rey Vega

Information-Theoretic Methods in Data Science

Chapter: Information Bottleneck and Representation Learning

Pablo Piantanida and Leonardo Rey Vega

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 792464".

Contents

1 Information Bottleneck and Representation Learning

A grand challenge in representation learning is the development of computational algorithms that learn the different explanatory factors of variation behind high-dimensional data. Representation models (usually referred to as encoders) are often determined for optimizing performance on training data when the real objective is to generalize well to other (unseen) data. The first part of this chapter is devoted to provide an overview of and introduction to fundamental concepts in statistical learning theory and the Information Bottleneck principle. It serves as a mathematical basis for the technical results given in the second part, in which an upper bound to the *generalization gap* corresponding to the *cross-entropy risk* is given. When this penalty term times a suitable multiplier and the *cross entropy empirical risk* are minimized jointly, the problem is equivalent to optimizing the Information Bottleneck objective with respect to the empirical data distribution. This result provides an interesting connection between mutual information and generalization, and helps to explain why noise injection during the training phase can improve the generalization ability of encoder models and enforce invariances in the resulting representations.

1.1 Introduction and Overview

Information theory aims to characterize the fundamental limits for data compression, communication, and storage. Although the coding techniques used to prove these fundamental limits are impractical, these provide valuable insight, highlighting key properties of good codes and leading to designs approaching the theoretical optimum (e.g., turbo codes, ZIP and JPEG compression algorithms). On the other hand, statistical models and machine learning are used to acquire knowledge from data. Models identify relationships between variables that allow making predictions and assessing their accuracy. A good choice of data representation is paramount for performing large-scale data processing in a computationally efficient and statistically meaningful manner [1], allowing to decrease storage, or to reduce inter-node communication if the data is distributed.

Shannon's abstraction of information merits careful study [2]. While a layman might think that the problem of communication is to convey meaning, Shannon clarified that *"the fundamental problem of communication is that of reproducing*

at one point a message selected at another point." Shannon further argued that the meaning of a message is subjective, i.e., dependent on the observer, and irrelevant to the engineering problem of communication. However, what does matter for the theory of communication is finding suitable representations for given data. In source coding, for example, one generally aims at distilling the relevant information from the data by removing unnecessary redundancies. This can be cast in information-theoretic terms, as higher redundancy makes data more predictable and lowers its information content.

In the context of learning [3, 4], we propose to distinguish these two rather different aspects of data: *information* and *knowledge*. *Information* contained in data is unpredictable and random, while additional structure and redundancy in the data stream constitutes *knowledge* about the data generation process, which a learner must acquire. Indeed, according to connectionist models [5], the redundancy contained within messages enables the brain to build up its cognitive maps and the statistical regularities in these messages are being used for this purpose. Hence, this *knowledge*, provided by redundancy [6, 7] in the data, must be what drives unsupervised learning. While information theory is a unique success story, from its birth, it discarded *knowledge* as being irrelevant to the engineering problem of communication. However, knowledge is recognized as being a critical –almost central– component of representation learning. The present monograph provides an information-theoretic treatment of this problem.

Knowledge representation. The data deluge of recent decades leads to new expectations for scientific discoveries from massive data. While mankind is drowning in data, a significant part of it is unstructured and 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. In many ways, *deep neural networks* has turned out to be very good at discovering structures in high-dimensional data and has dramatically improved the state of the art in several pattern recognition tasks [8]. The global learning task is decomposed into a hierarchy of layers with nonlinear processing, having great success not only due to their ability to fit different types of datasets but also to generalize incredible well. The representational capabilities of neural networks [9] have drawn significant interest from the machine learning community. These networks seem to be able to learn multi-level abstractions with capability to harness unlabeled data, multi-task learning, and multiple inputs, while learning from distributed and hierarchical data, to represent context at multiple levels.

The actual goal of representation learning is neither accurate estimation of model parameters [10] nor compact representation of the data itself [11, 12]; rather, we are mostly interested in the generalization capabilities, meaning its ability to successfully apply rules extracted from previously seen data to characterize unseen data. According to the statistical learning theory [13], models with many parameters tend to *overfit* by representing the learned data too accu-

rately, therefore diminishing their ability to generalize to unseen data. In order to reduce this 'generalization gap', i.e., the difference between 'training error' and 'test error' (a measure of how well the learner has learned), several regularization methods where proposed in the literature [13]. A recent breakthrough in this area has been the development of *Dropout* [14] for training deep neural networks. This consists in randomly dropping units during training to prevent their co-adaptation, including some information-based regularization [15] that yields a slightly more general form of the variational auto-encoder [16].

Why is that we succeed in learning high-dimensional representations? Recently there has been much interest in understanding the importance of implicit regularization. Numerical experiments in [17] demonstrate that network size may not be the main form of capacity control for deep neural networks and hence, some other unknown form of capacity control plays a central role in learning multi-layer feed-forward networks. From a theoretical perspective regularization seems to be an indispensable component in order to improve the final misclassification probability while convincing experiments support the idea that the absence of all regularization does not necessarily induces poor generalization gap. Possible explanations were approached via rate-distortion theory [18,19] by exploring heuristic connections with the celebrate Information Bottleneck principle [20]. Along the same line of work, [21,22] have proven bounds showing that the square root of the mutual information between the training inputs and the parameters inferred from the training algorithm provides a concise bound on the generalization gap. These bounds crucially depend on the Markov operator that maps the training set into the network parameters and whose characterization could not be an easy task. Similarly, [23] explored how the use of an Information Bottleneck objective on the network parameters (and not on the representations) may help to avoid overfitting while enforcing invariant representations.

The interplay between information and complexity. The goal of data representation may be cast as trying to find regularity in the data. Regularity may be identified with "ability to compress" by viewing representation learning as lossy data compression: it tells us that, for a given set of encoder models and dataset, we should try to find the encoder or combination of encoders that compresses the data most. In this sense, we may speak of the information complexity of a structure, meaning the minimum amount of information (number of bits) we need to store enough information about the structure that allows us its reconstruction. The central result in this chapter states that good representation models should squeeze out as much regularity as possible from the given data. In other words, representations are expected to distill the *meaningful information* present in the data, i.e., to separate structure as seeing the regularity from noise, interpreted as the *accidental information*.

Structure of this chapter. This chapter can be read without any prior knowledge of information theory and statistical learning theory. In the first part, the basic learning framework for analysis is developed and an accessible overview of basic concepts in statistical learning theory and the information bottleneck

principle are presented. The second part introduces an upper bound to the generalization gap corresponding to the cross-entropy loss and show that when this penalty term times a suitable multiplier and plus the cross entropy empirical risk are minimized jointly, the problem is equivalent to optimizing the Information Bottleneck objective with respect to the empirical data distribution. The notion of information complexity is introduced and intuitions behind it are developed.

1.2 Representation and Statistical Learning

We introduce the framework by which leaning from examples is to be studied. We develop precise notions of risk and the generalization gap, and discuss the mathematical factors upon which these depend.

1.2.1 Basic Definitions

In this monograph we are concerned with the problem of pattern classification which is about predicting the unknown class of an example or observation. An example can be modelled as an information source $X \in \mathcal{X}$ presented to the learner about a target concept $Y \in \mathcal{Y}$ (the concept class). In our model we simply assume $(\mathcal{X}, \mathcal{Y})$ are abstract discrete spaces equipped with a σ-algebra. In the problem of pattern classification, one searches for a function $c : \mathcal{X} \longrightarrow \mathcal{Y}$ which represents one's guess of Y given X. Although there is much to say about statistical learning, this section does not cover extensively the field (an overview can be found in [13]). Besides this, we limit ourselves to describing the key ideas in a simple way, often sacrificing generality.

DEFINITION 1.1 (Misclassification probability) An $|\mathcal{Y}|$-ary classifier is defined by a (possibly stochastic) decision rule $Q_{\hat{Y}|X} : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$, where $\hat{Y} \in \mathcal{Y}$ denotes the random variable associated to the classifier output and X is the information source. The probability of misclassification of a rule $Q_{\hat{Y}|X}$ with respect to a data distribution P_{XY} is given by:

$$P_{\mathcal{E}}\big(Q_{\hat{Y}|X}\big) \coloneqq 1 - \mathbb{E}_{P_{XY}}\left[Q_{\hat{Y}|X}(Y|X)\right]. \tag{1.1}$$

Minimizing over all possible classifiers $Q_{\hat{Y}|X}$ gives the smallest average probability of misclassification. An optimum classifier $c^{\star}(\cdot)$ chooses the hypothesis $\hat{y} \in \mathcal{Y}$ with largest posterior probability $P_{Y|X}$ given the observation x, that is the *Maximum a Posteriori (MAP)* decision. The MAP test that breaks ties randomly with equal probability is given by[1]

$$Q^{\mathrm{MAP}}_{\hat{Y}|X}(y|x) \coloneqq \begin{cases} \dfrac{1}{|\mathcal{B}(x)|}, & \text{if } y \in \mathcal{B}(x) \\ 0, & \text{otherwise} \end{cases} \tag{1.2}$$

[1] In general, the optimum classifier given in (1.2) is not unique. Any conditional pmf with support in $\mathcal{B}(x)$ for each $x \in \mathcal{X}$ will be equally good.

where the set $\mathcal{B}(x)$ is defined as: $\mathcal{B}(x) \coloneqq \left\{ y \in \mathcal{Y} : \; P_{Y|X}(y|x) = \max_{y' \in \mathcal{Y}} P_{Y|X}(y'|x) \right\}$. This classification rule is called *Bayes decision rule*. Bayes decision rule is optimal in the sense that no other decision rule has a smaller probability of misclassification. It is straightforward to obtain the following lemma:

LEMMA 1.2 (Bayes error) *The misclassification error rate of the Bayes decision rule is given by*

$$P_{\mathcal{E}}\left(Q^{MAP}_{\hat{Y}|X}\right) = 1 - \mathbb{E}_{P_X}\left[\max_{y' \in \mathcal{Y}} P_{Y|X}(y'|X)\right]. \tag{1.3}$$

Finding the Bayes decision rule requires knowledge of the underlying distribution P_{XY}, but typically in applications these distributions are not known. In fact, even a parametric form or an approximation to the true distribution is unknown. In this case, the *learner* tries to overcome the lack of knowledge by resorting to labeled examples. In addition, the probability of misclassification using the labeled examples has the particularity that it is mathematically hard to solve for the optimal decision rule. As a consequence, it is common to work with a surrogate (information measure) given by the *average logarithmic loss or cross-entropy loss*. This loss is used when a probabilistic interpretation of the scores is desired by measuring the dissimilarity between the true label distribution $P_{Y|X}$ and the predicted label distribution $Q_{\hat{Y}|X}$, and is defined below.

LEMMA 1.3 (Surrogate based on the average logarithmic loss) *A natural surrogate for the probability of misclassification $P_{\mathcal{E}}(Q_{\hat{Y}|X})$ corresponding to a classifier $Q_{\hat{Y}|X}$ is given by the average logarithmic loss $\mathbb{E}_{P_{XY}}\left[-\log Q_{\hat{Y}|X}(Y|X)\right]$ which satisfies:*

$$P_{\mathcal{E}}(Q_{\hat{Y}|X}) \leq 1 - \exp\left(-\mathbb{E}_{P_{XY}}\left[-\log Q_{\hat{Y}|X}(Y|X)\right]\right). \tag{1.4}$$

A lower-bound for the average logarithmic loss can be computed as:

$$\mathbb{E}_{P_{XY}}\left[-\log Q_{\hat{Y}|X}(Y|X)\right] \geq H(P_{Y|X}|P_X). \tag{1.5}$$

The average logarithmic loss can provide an effective and better behaved surrogate for the particular problem of minimizing the probability of misclassification [9]. Evidently, the optimal decision rule for the average logarithmic loss is $Q_{\hat{Y}|X} \equiv P_{Y|X}$. This does not match in general with the optimal decision rule for the probability of misclassification $Q^{\text{MAP}}_{\hat{Y}|X}$ in expression (1.2). Although the average logarithmic loss may induce an irreducible gap with respect to the probability of misclassification, it is clear that when the true $P_{Y|X}$ concentrates around a particular value $y(x)$ for each $x \in \mathcal{X}$ (which is necessary for a statistical model $P_{Y|X}$ to induce low probability of misclassification) this gap could be significantly reduced.

1.2.2 Learning Data Representations

We will concern ourselves with learning representation models (randomized encoders) and self-classifiers (randomized decoders) from labeled examples. In other words, learning target probability distributions which are assumed to belong to some class of distributions. The motivation behind this paradigm relies on a view of the brain as an information processor that in solving certain problems (e.g. object recognition) builds a series of internal representations starting with the sensory (external) input from which it computes a function (e.g. detecting the orientations of edges in an image or learning to recognize individual faces).

The problem of finding a good classifier can be divided into that of simultaneously finding a (possibly randomized) encoder $Q_{U|X} : \mathcal{X} \to \mathcal{P}(\mathcal{U})$ that maps raw data to a representation, possibly living in a higher-dimensional (feature) space $\mathcal{U}$ and a soft-decoder $Q_{\hat{Y}|U} : \mathcal{U} \to \mathcal{P}(\mathcal{Y})$ which maps the representation to a probability distribution on the label space $\mathcal{Y}$. Although these mappings induce an equivalent classifier:

$$Q_{\hat{Y}|X}(y|x) = \sum_{u \in \mathcal{U}} Q_{U|X}(u|x) Q_{\hat{Y}|U}(y|u), \tag{1.6}$$

the computation of the later expression requires marginalizing out $u \in \mathcal{U}$ which is in general computationally hard due to the exponential number of atoms involved in the representations. A variational upper bound is commonly used to rewrite this intractable problem into:

$$\mathbb{E}_{P_{XY}} \left[-\log Q_{\hat{Y}|X}(Y|X) \right] \leq \mathbb{E}_{P_{XY}} \mathbb{E}_{Q_{U|X}} \left[-\log Q_{\hat{Y}|U}(Y|U) \right], \tag{1.7}$$

which simply follows by applying Jensen inequality [24]. This bound induces the well-known *cross-entropy risk* defined below:

DEFINITION 1.4 (Cross-entropy loss and risk) Given two randomized mappings $Q_{U|X} : \mathcal{X} \to \mathcal{P}(\mathcal{U})$ and $Q_{\hat{Y}|U} : \mathcal{U} \to \mathcal{P}(\mathcal{Y})$, we define the average (over representations) *cross-entropy loss* as:

$$\ell\big(Q_{U|X}(\cdot|x), Q_{\hat{Y}|U}(y|\cdot)\big) \coloneqq \big\langle Q_{U|X}(\cdot|x), -\log Q_{\hat{Y}|U}(y|\cdot) \big\rangle \tag{1.8}$$

$$= -\sum_{u \in \mathcal{U}} Q_{U|X}(u|x) \log Q_{\hat{Y}|U}(y|u). \tag{1.9}$$

We measure the expected performance of $(Q_{U|X}, Q_{\hat{Y}|U})$ via the *risk* function:

$$(Q_{U|X}, Q_{\hat{Y}|U}) \mapsto \mathcal{L}(Q_{U|X}, Q_{\hat{Y}|U}) \coloneqq \mathbb{E}_{P_{XY}} \big[\ell\big(Q_{U|X}(\cdot|X), Q_{\hat{Y}|U}(Y|\cdot)\big) \big]. \tag{1.10}$$

In addition to the points noted earlier, another crucial component of knowledge representation is the use of deep representations. Formally speaking, we consider K-th randomized encoders $\{Q_{U_k|U_{k-1}}\}_{k=1}^{K}$ with $U_0 \equiv X$ instead of one randomized encoder $Q_{U|X}$. Although this appears at first to be more general, it can be casted using the one-randomized encoder formulation induced by the marginal distribution that relates the input and the output layer of the network. Therefore any result for the one-layer formulation immediately implies a result

for the K-th layer formulation and for this reason we shall focus on the one-layer case without loss of generality.

LEMMA 1.5 (Optimal decoders) *The minimum cross-entropy loss risk satisfies:*

$$\inf_{Q_{\hat{Y}|U}:\mathcal{U}\to\mathcal{P}(\mathcal{Y})} \mathcal{L}(Q_{\hat{Y}|U}, Q_{U|X}) = H(Q_{Y|U}|Q_U), \tag{1.11}$$

where

$$Q_{Y|U}(y|u) = \frac{\sum\limits_{x\in\mathcal{X}} Q_{U|X}(u|x)P_{XY}(x,y)}{\sum\limits_{x\in\mathcal{X}} Q_{U|X}(u|x)P_X(x)}. \tag{1.12}$$

Proof The proof follows from the positivity of the relative entropy by noticing that $\mathcal{L}(Q_{U|X}, Q_{\hat{Y}|U}) = D\left(Q_{Y|U}\middle\|Q_{\hat{Y}|U}\middle|Q_U\right) + H(Q_{Y|U}|Q_U)$. □

The associated risk to the optimal decoder is:

$$\mathcal{L}(Q_{U|X}, Q_{Y|U}) \coloneqq \mathbb{E}_{P_{XY}}\left[-\sum_{u\in\mathcal{U}} Q_{U|X}(u|x)\log Q_{Y|U}(Y|U)\right], \tag{1.13}$$

which is only a function of the encoder model $Q_{U|X}$. However, the optimal decoder cannot be determined since P_{XY} is unknown.

The learner's goal is to select $Q_{U|X}$ and $Q_{\hat{Y}|U}$ by minimizing the risk (1.10). However, since P_{XY} is unknown the learner cannot directly measure the risk and it is common to measure the agreement of a pair of candidates with a finite training dataset based on the *empirical risk*.

DEFINITION 1.6 (Empirical risk) Let $\hat{P}_{XY}$ denote the empirical distribution through the training dataset $\mathcal{S}_n \coloneqq \{(x_1, y_1), \ldots, (x_n, y_n)\}$. The empirical risk is:

$$\mathcal{L}_{\text{emp}}(Q_{U|X}, Q_{\hat{Y}|U}) \coloneqq \mathbb{E}_{\hat{P}_{XY}}\left[\ell\big(Q_{U|X}(\cdot|X), Q_{\hat{Y}|U}(Y|\cdot)\big)\right] \tag{1.14}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \ell\big(Q_{U|X}(\cdot|x_i), Q_{\hat{Y}|U}(y_i|\cdot)\big). \tag{1.15}$$

LEMMA 1.7 (Optimality of empirical decoders) *Given a randomized encoder $Q_{U|X} : \mathcal{X} \to \mathcal{P}(\mathcal{U})$, define the empirical decoder with respect to the empirical distribution $\hat{P}_{XY}$ as:*

$$\hat{Q}_{Y|U}(y|u) \coloneqq \frac{\sum\limits_{x\in\mathcal{X}} Q_{U|X}(u|x)\hat{P}_{XY}(x,y)}{\sum\limits_{x\in\mathcal{X}} Q_{U|X}(u|x)\hat{P}_X(x)}. \tag{1.16}$$

Then, the risk can be lower bounded uniformly over $Q_{\hat{Y}|U} : \mathcal{U} \to \mathcal{P}(\mathcal{Y})$ as:

$$\mathcal{L}_{emp}(Q_{U|X}, Q_{\hat{Y}|U}) \geq \mathcal{L}_{emp}(Q_{U|X}, \hat{Q}_{Y|U}), \tag{1.17}$$

where equality holds provided that $Q_{\hat{Y}|U} \equiv \hat{Q}_{Y|U}$, i.e., the optimal decoder is computed from the encoder and the empirical distribution as done in (1.16).

Proof The inequality follows along the lines of Lemma (1.5) by noticing that $\mathcal{L}_{\text{emp}}(Q_{U|X}, Q_{\hat{Y}|U}) = D\left(\hat{Q}_{Y|U} \| Q_{\hat{Y}|U} | \hat{Q}_U\right) + \mathcal{L}_{\text{emp}}(\hat{Q}_{Y|U}, Q_{U|X})$. Finally, the non-negativity of relative conditional entropy completes the proof. □

Since the empirical risk is evaluated on finite samples, its evaluation may be sensitive to sampling (noise) error and thus giving rise to the issue of generalization. It can be argued, that a key component of learning is not just the development of a representation model on the basis of finite training dataset, but its use in order to generalize to unseen data. Clearly successful generalization necessitates the closeness (in some sense) of the selected representation and decoder models. Therefore, successful representation learning would involve successful generalization. This chapter deals with the information complexity of successful generalization. The generalization gap defined below is a measure of how an algorithm could perform on new data, i.e., data that is not available during the training phase. In the light of Lemmas 1.5 and 1.7, we will restrict our analysis to encoders only and assume that the optimal empirical decoder has been selected, i.e., $Q_{\hat{Y}|U} \equiv \hat{Q}_{Y|U}$ in both the empirical (1.14) and the true (1.10) risks. This is reasonable given the fact that the true P_{XY} is not known and the only decoder that can be implemented in practice is the empirical one.

DEFINITION 1.8 (Generalization gap) Given a stochastic mapping $Q_{U|X} : \mathcal{X} \to \mathcal{P}(\mathcal{U})$, the *generalization gap* is defined as:

$$(Q_{U|X}, \mathcal{S}_n) \mapsto \mathcal{E}_{\text{gap}}(Q_{U|X}, \mathcal{S}_n) \coloneqq \left| \mathcal{L}_{\text{emp}}(Q_{U|X}, \hat{Q}_{Y|U}) - \mathcal{L}(Q_{U|X}, \hat{Q}_{Y|U}) \right|, \tag{1.18}$$

which represents the error incurred by the selected $Q_{U|X}$ when the rule $\mathcal{L}_{\text{emp}}(Q_{U|X}, \hat{Q}_{Y|U})$ is used instead of the true risk $\mathcal{L}(Q_{U|X}, \hat{Q}_{Y|U})$.

1.2.3 Optimizing on Restricted Classes of Randomized Encoders

We have already introduced the notions of representation and inference models and risk functions from which these candidates are chosen. Another related question of interest is: how do we define the encoder class? A simple approach is to model classes in a parametric fashion. We first introduce the Bayes risk and then the restricted classes of randomized encoders and decoders.

DEFINITION 1.9 (Bayes risk) The minimum cross-entropy risk over all possible candidates is called the *Bayes risk* and will be denoted by $\mathcal{L}^\star$. In this case,

$$\mathcal{L}^\star \coloneqq \inf_{Q_{U|X}: \mathcal{X} \to \mathcal{P}(\mathcal{U})} \mathcal{L}\big(Q_{U|X}, \hat{Q}_{Y|U}\big) = H(P_{Y|X}|P_X). \tag{1.19}$$

DEFINITION 1.10 (Learning model) The encoder functions are defined by $f_{\boldsymbol{\theta}} : \mathcal{X}^d \times \mathcal{Z} \longrightarrow \mathcal{U}_{\boldsymbol{\theta}}^m$, where $\mathcal{X}$ is the finite input alphabet with cardinality $|\mathcal{X}|$ and d is a positive integer, $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^{d_\Theta}$ denotes the unknown parameters to be optimized, Z is a random variable taking values on a finite alphabet $\mathcal{Z}$ with

probability P_Z whose role is to randomize encoders and $\mathcal{U}_{\boldsymbol{\theta}} \subset [0,1]$ is the alphabet corresponding to the hidden representation which satisfies $|\mathcal{U}_{\boldsymbol{\theta}}| \leq |\mathcal{X}|^d \cdot |\mathcal{Z}|$. For notational convenience, we let $\mathcal{X} \equiv \mathcal{X}^d$ and $\mathcal{U} \equiv \mathcal{U}_\theta^m$ and denote this class as:

$$\mathcal{F} \coloneqq \left\{ Q_{U|X}(u|x) = \mathbb{E}_{P_Z}\left[\mathbb{1}[u = f_{\boldsymbol{\theta}}(x,Z)]\right] \ : \ \boldsymbol{\theta} \in \Theta \right\}.$$

It is clear that for every $\boldsymbol{\theta}$, $\boldsymbol{\theta} \mapsto Q_{U|X} \in \mathcal{F}$ induces a randomized encoder.

In order to simplify subsequent analysis we will assume the following conditions over the possible data pmf and over the family $\mathcal{F}$ of encoders:

DEFINITION 1.11 (Restricted model class) We assume that alphabets $\mathcal{X}$, $\mathcal{Y}$ are of arbitrary large size but finite. Furthermore, there exists $\eta > 0$ such that the unknown data generating distribution P_{XY} satisfies $P_X(x_{\min}) \coloneqq \min_{x\in\mathcal{X}} P_X(x) \geq \eta$ and $P_Y(y_{\min}) \coloneqq \min_{y\in\mathcal{Y}} P_Y(y) \geq \eta$.

DEFINITION 1.12 (Empirical risk minimization) The methodology of empirical risk minimization is one of the most straight-forward, yet it is usually efficient provided that the chosen model class $\mathcal{F}$ is restricted [25]. The learner chooses a pair $\hat{Q}^\star_{U|X} \in \mathcal{F}$ that minimizes the empirical risk:

$$\mathcal{L}_{\text{emp}}\big(\hat{Q}^\star_{U|X}, \hat{Q}^\star_{Y|U}\big) \leq \mathcal{L}_{\text{emp}}(Q_{U|X}, \hat{Q}_{Y|U}), \quad \text{for all } Q_{U|X} \in \mathcal{F}. \tag{1.20}$$

Moreover, it is possible to minimize a surrogate of the true risk:

$$\mathcal{L}(Q_{U|X}, \hat{Q}_{Y|U}) \leq \mathcal{L}_{\text{emp}}(Q_{U|X}, \hat{Q}_{Y|U}) + \mathcal{E}_{\text{gap}}(Q_{U|X}, \mathcal{S}_n), \tag{1.21}$$

which depends on the empirical risk and the so-called generalization gap, respectively. Expression (1.21) states that an adequate selection of the encoder should be performed in order to minimize the empirical risk and the generalization gap simultaneously. It is reasonable to expect that the assumption that the optimal encoder achieving the minimal risk in (1.19) do not belong to our restricted class of models $\mathcal{F}$, so the learner may want to enlarge the model classes $\mathcal{F}$ as much as possible. However, this could induce a larger value of the generalization gap, which could lead to a tradeoff between these two fundamental quantities.

DEFINITION 1.13 (Approximation and estimation error) The suboptimality of the model class $\mathcal{F}$ is measured in terms of the approximation error:

$$\mathcal{E}_{\text{app}}(\mathcal{F}) \coloneqq \inf_{Q_{U|X}\in\mathcal{F}} \mathcal{L}(Q_{U|X}, \hat{Q}_{Y|U}) - \mathcal{L}^\star. \tag{1.22}$$

The induced risk of the selected pair of encoder is given by

$$\mathcal{E}_{\text{est}}\big(\mathcal{F}, \hat{Q}^\star_{U|X}, \hat{Q}^\star_{Y|U}\big) \coloneqq \mathcal{L}\big(\hat{Q}^\star_{U|X}, \hat{Q}^\star_{Y|U}\big) - \inf_{Q_{U|X}\in\mathcal{F}} \mathcal{L}(Q_{U|X}, \hat{Q}_{Y|U}), \tag{1.23}$$

where $\hat{Q}^\star_{U|X}$ denotes the minimizer of expression (1.20).

DEFINITION 1.14 (Excess risk) The *excess risk* of the algorithm (1.20) selecting an optimal pair $(\hat{Q}^\star_{Y|U}, \hat{Q}^\star_{U|X})$ can be decomposed as:

$$\begin{aligned}\mathcal{E}_{\text{exc}}\big(\mathcal{F}, \hat{Q}^\star_{U|X}, \hat{Q}^\star_{Y|U}\big) &\coloneqq \mathbb{E}\big[\mathcal{L}\big(\hat{Q}^\star_{U|X}, \hat{Q}^\star_{Y|U}\big)\big] - \mathcal{L}^\star \\ &= \mathcal{E}_{\text{app}}(\mathcal{F}) + \mathbb{E}\big[\mathcal{E}_{\text{est}}\big(\mathcal{F}, \hat{Q}^\star_{U|X}, \hat{Q}^\star_{Y|U}\big)\big],\end{aligned}$$

where the expectation is taken with respect to the random choice $\mathcal{S}_n$ of dataset which induces the optimal pair $(\hat{Q}^\star_{Y|U}, \hat{Q}^\star_{U|X})$.

The *approximation error* $\mathcal{E}_{\text{app}}(\mathcal{F})$ measures how closely encoders in the model class $\mathcal{F}$ can approximate the optimal solution $\mathcal{L}^\star$. On the other hand the *estimation error* $\mathcal{E}_{\text{est}}(\mathcal{F}, \hat{Q}^\star_{U|X}, \hat{Q}^\star_{Y|U})$ measures the effect of minimizing the empirical risk instead of the true risk, caused by the finite size of the training data. The estimation error is determined by the number of training samples and by the complexity of the model class, i.e., large models have smaller approximation error but lead to higher estimation errors, and it is also related to the generalization error [25]. However, for the sake of simplicity, in this chapter we restrict our attention only to the generalization gap.

1.3 Information-Theoretic Principles and Information Bottleneck

1.3.1 Lossy Source Coding

The problem of source coding is jointly with the channel coding one, the two more important and relevant problems in information theory [24, 26]. In the source coding problem, we faced the fundamental question: how to represent in the most compact way a given stochastic source such that we can be able to reconstruct, with a given level of fidelity, the original source. Shannon was the first to formalize and solve completely this problem [2, 27] in the asymptotic regime[2], establishing the optimal trade-off between the compactness of the representations and the level of fidelity in the reconstruction. In the *lossless source coding* problem, the level of required fidelity is maximal: it is desired to have a short-length representation which can be used to reconstruct almost exactly the original source. According to the more general *lossy source coding* setup, we look for more compact representations of the original source by dropping the requirement of almost-exact reconstruction. The level of fidelity required is measured by using a predefined *distortion measure* which is a an essential part of the problem. Interestingly enough, the problem of lossy source coding can be solved completely for any well-defined distortion measure. Let us mathematically formulate this problem:

[2] The asymptotic regime considers that the number of realizations of the stochastic source to be compressed tends to infinity. Although this could be questionable in practice, the asymptotic problem reflects accurately the important trade-offs of the problem. In this presentation, our focus will be the in the asymptotic problem originally solved by Shannon.

DEFINITION 1.15 (Lossy source coding problem) Consider a discrete and finite alphabet $\mathcal{X}$ and a stochastic source X, which generates identically and independently distributed samples according to $P_X \in \mathcal{P}(\mathcal{X})$. Consider an alternative alphabet $\hat{\mathcal{X}}$ and a distortion function $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}_{\geq 0}$. Consider also a realization X^n of the source, an *encoder* function $f_n : \mathcal{X}^n \to \{1, \dots M_n\}$ where $M_n \in \mathbb{N}$ and a *decoder* function $g_n : \{1, \dots, M_n\} \to \hat{\mathcal{X}}^n$. We say that $\mathcal{C}_n = (f_n, g_n)$ is an n-code for X and $d(\cdot;\cdot)$.

DEFINITION 1.16 (Achievable rate and fidelity) A pair (R, D) is said to be achievable if for every $\epsilon > 0$, there exist $n \geq 1$ and an n-code $\mathcal{C}_n$ such that:

$$\frac{\log M_n}{n} \leq R + \epsilon, \tag{1.24}$$

$$\mathbb{E}_{P_X^n}\left[\bar{d}\left(X^n; g_n\left(f_n(X^n)\right)\right)\right] \leq D + \epsilon, \tag{1.25}$$

where $\bar{d}(X^n; \hat{X}^n) \equiv \frac{1}{n}\sum_{i=1}^n d(x_i; \hat{x}_i)$.

The set of all achievable pairs (R, D) contains the complete characterization of all the possible trade-offs between the *rate* R (which quantifies the level of compression of the source X measuring the necessary number of bits per symbol) and the *distortion* D (which quantifies the average fidelity level per symbol in the reconstruction using the distortion function $d(\cdot;\cdot)$ symbol by symbol). An equivalent characterization of the set of achievable pairs (R, D) is given by the *rate-distortion* function defined by:

$$R(D) = \inf\{R : (R, D) \text{ is achievable}\}. \tag{1.26}$$

It is the great achievement of Shannon [27] to have obtained the following result:

THEOREM 1.17 (Rate-distortion function) *The rate-distortion function for source X with reconstruction alphabet $\hat{\mathcal{X}}$ and with distortion function $d(\cdot;\cdot)$ is given by:*

$$\mathcal{R}_{X,d}(D) = \inf_{\substack{P_{\hat{X}|X}:\, \mathcal{X} \to \mathcal{P}(\hat{\mathcal{X}}) \\ \mathbb{E}_{P_{\hat{X}X}}[d(X;\hat{X})] \leq D}} I\big(P_X; P_{\hat{X}|X}\big). \tag{1.27}$$

This function depends on solely on the distribution P_X and the distortion function $d(\cdot;\cdot)$ and contains the exact trade-off between compression and fidelity that can be expected for the particular source and distortion function. It is easy to establish that this function is positive, non-increasing in D and convex. Moreover, there exists $D > 0$ such that $\mathcal{R}_{X,d}(D)$ is finite and we denote the minimum of such values of D by $D_{\min}$ with $R_{\max} \coloneqq \lim_{D \to D_{\min}+} \mathcal{R}_{X,d}(D)$. Although $\mathcal{R}_{X,d}(D)$ could be hard to compute in closed form for a particular P_X and $d(\cdot;\cdot)$, the problem in (1.27) is a convex optimization one, for which there exist efficient numerical techniques. However, several important cases admit closed form expressions as the Gaussian case with quadratic distortion[3] [24].

[3] Although the Gaussian case does not correspond to a finite cardinality set $\mathcal{X}$, the result in (1.27) can be easily extended to such case using quantization arguments.

Other important function related with the rate-distortion function is the *distortion-rate* function. This function can be defined independently from the rate-distortion function and directly from information-theoretic principles. Intuitively, this function is the infimum value of the distortion D as function of the rate R for all (R, D) achievable pairs. We will define it directly from the rate-distortion function:

$$\mathcal{R}_{X,d}^{-1}(I) \coloneqq \inf \left\{ D \in \mathbb{R}_{\geq 0} : \mathcal{R}_{X,d}(D) \leq I \right\}. \tag{1.28}$$

It is not hard to show the following[4]:

LEMMA 1.18 *Consider the distortion-rate function defined according to (1.28). This function is positive, non-increasing in I and convex.*

Proof Follow easily from definition (1.28). □

Besides their obvious importance in the problem of source coding, the definitions of the rate-distortion and distortion-rate functions will be useful for the problem of learning as presented in the previous section. They will permit on establish connections between the misclassification probability, the cross-entropy and the mutual information between the input X and the output of the encoder $Q_{U|X}$. These connections will be conceptually important for the rest of the chapter, at least from a qualitative point of view.

1.3.2 Misclassification Probability and Cross-Entropy Loss

It is easy to show that the proposed learning framework can be set up as a lossy-source coding problem. This formulation, however, it is not an operational one as was the case for the information-theoretic one presented in Definitions 1.15 and 1.16. The reason for this comes form the fact that for our learning framework we do not have the same type of scaling with n as in the source coding problem in information theory. While in the typical source coding problem, encoders and decoders act upon the entire sequence of observed samples $x^n = (x_1, \ldots, x_n)$, in the learning framework, the encoder $Q_{U|X}$ acts on sample-by-sample basis. Nevertheless, the definition of the rate-distortion (w.r.t. distortion-rate) function is relevant for the learning framework as well, provided that we avoid any operational interpretation and concentrate on their strictly mathematical meaning.

Consider alphabets $\mathcal{U}$, $\mathcal{X}$ and $\mathcal{Y}$, corresponding to the descriptions generated by the encoder $Q_{U|X}$ and to the examples and their corresponding labels. From

[4] It is worth to mention that by using $\mathcal{R}_{X,d}^{-1}(I)$ we are abusing notation. This is because in general is not true that $\mathcal{R}_{X,d}(D)$ is injective for every $D \geq 0$. However, when $I \in [R_{\min}, R_{\max})$ with $R_{\min} \coloneqq \mathcal{R}_{X,d}(D_{\max})$ and $D_{\max} \coloneqq \min_{\hat{x} \in \hat{\mathcal{X}}} \mathbb{E}_{P_X}\left[d(X; \hat{x})\right]$, under some very mild conditions on P_X and $d(\cdot;\cdot)$, $\mathcal{R}_{X,d}^{-1}(I)$ is the true inverse of $\mathcal{R}_{X,d}(D)$, which is guaranteed to be injective in the interval $D \in (D_{\min}, D_{\max}]$.

(1.1) and (1.6), we can write the misclassification probability as:

$$\begin{aligned} P_{\mathcal{E}}\big(Q_{U|X}, Q_{\hat{Y}|U}\big) &= 1 - \mathbb{E}_{P_{XY}}\left[\sum_{u\in\mathcal{U}} Q_{U|X}(u|X) Q_{\hat{Y}|U}(Y|u)\right] \\ &= 1 - \mathbb{E}_{P_Y}\left[\sum_{u\in\mathcal{U}} Q_{\hat{Y}|U}(Y|u)\mathbb{E}_{P_{X|Y}}\left[Q_{U|X}(u|X)\right]\right] \\ &= 1 - \mathbb{E}_{P_Y}\left[\sum_{u\in\mathcal{U}} Q_{\hat{Y}|U}(Y|u) Q_{U|Y}(u|Y)\right] \\ &= \mathbb{E}_{P_{UY}}\left[1 - Q_{\hat{Y}|U}(Y|U)\right]. \end{aligned} \tag{1.29}$$

From the above derivation, we can set a distortion measure: $d(u;y) \coloneqq 1 - Q_{\hat{Y}|U}(y|u)$. In this way, the probability of misclassification can be written as an average over the outcomes of Y (taking as the source) and U (taking as the reconstruction) of the distortion measure: $1 - Q_{\hat{Y}|U}(y|u)$. In this manner, we can consider the following rate-distortion function:

$$\mathcal{R}_{Y,Q_{\hat{Y}|U}}(D) \coloneqq \inf_{\substack{P_{U|Y}:\mathcal{Y}\to\mathcal{P}(\mathcal{U}) \\ \mathbb{E}_{P_{UY}}[1-Q_{\hat{Y}|U}(Y|U)]\le D}} I\big(P_Y; P_{U|Y}\big), \tag{1.30}$$

which provides a connection between the misclassification probability and mutual information $I\big(P_Y; P_{U|Y}\big)$.

From this formulation we are able to obtain the following lemma, which provides an upper and a lower bound on the probability of misclassification via the distortion-rate function and the cross-entropy loss.

LEMMA 1.19 (Probability of misclassification and cross-entropy loss) *The probability of misclassification $P_{\mathcal{E}}(Q_{\hat{Y}|U}, Q_{U|X})$ induced by a randomized encoder $Q_{U|X} : \mathcal{X} \to \mathcal{P}(\mathcal{U})$ and decoder $Q_{\hat{Y}|U} : \mathcal{U} \to \mathcal{P}(\mathcal{Y})$ is bounded by*

$$\begin{aligned} \mathcal{R}^{-1}_{Y,Q_{\hat{Y}|U}}\left(I(P_X; Q_{U|X})\right) &\le \mathcal{R}^{-1}_{Y,Q_{\hat{Y}|U}}\left(I(P_Y; Q_{U|Y})\right) && (1.31) \\ &\le P_{\mathcal{E}}(Q_{\hat{Y}|U}, Q_{U|X}) && (1.32) \\ &\le 1 - \exp\Big(-\mathcal{L}(Q_{\hat{Y}|U}, Q_{U|X})\Big), && (1.33) \end{aligned}$$

where $Q_{U|Y}(u|y) = \sum_{x\in\mathcal{X}} Q_{U|X}(u|x) P_{X|Y}(x|y)$ for $(u,y) \in \mathcal{U}\times\mathcal{Y}$.

Proof The upper bound simply follows by using Jensen-Inequality [24] while the lower bound is a consequence of the definition of the rate-distortion and distortion-rate functions. The probability of misclassification corresponding to the classifier can be expressed by the expected distortion $\mathbb{E}_{P_{XY}Q_{U|X}}[d(Y,U)] = P_{\mathcal{E}}(Q_{\hat{Y}|U}, Q_{U|X})$ based on the fidelity function $d(y,u) \coloneqq 1 - Q_{\hat{Y}|U}(y|u)$ as showed in (1.29). Because of the Markov chain $Y \multimap X \multimap U$, we can use the data processing inequality [24] and the definition of the rate-distortion function, obtaining the

following bound for the classification error:

$$I(P_X;Q_{U|X}) \geq I(P_Y;Q_{U|Y}) \tag{1.34}$$

$$\geq \inf_{\substack{P_{\hat{U}|Y}:\,\mathcal{Y}\to\mathcal{P}(\mathcal{U}) \\ \mathbb{E}_{P_{\hat{U}Y}}[d(Y,\hat{U})]\leq \mathbb{E}_{P_{XY}Q_{U|X}}[d(Y,U)]}} I\big(P_Y;P_{\hat{U}|Y}\big) \tag{1.35}$$

$$= \mathcal{R}_{Y,Q_{\hat{Y}|U}}\big(P_{\mathcal{E}}(Q_{\hat{Y}|U},Q_{U|X})\big). \tag{1.36}$$

For $\mathbb{E}_{P_{XY}Q_{U|X}}[d(Y,U)]$, we can use the definition of $\mathcal{R}^{-1}_{Y,Q_{\hat{Y}|U}}(\cdot)$, and thus obtain from (1.34) the fundamental bound

$$\mathcal{R}^{-1}_{Y,Q_{\hat{Y}|U}}(I(P_X;Q_{U|X})) \leq \mathcal{R}^{-1}_{Y,Q_{\hat{Y}|U}}(I(P_Y;Q_{U|Y})) \leq P_{\mathcal{E}}(Q_{\hat{Y}|U},Q_{U|X}).$$

□

The lower bound in the above expression states that any limitation in terms of mutual information between raw data and its representation will bound from below the probability of misclassification while the upper bound shows that cross-entropy loss introduced in (1.10) can be used as a surrogate to optimize the probability of misclassification, as it was also pointed out in Lemma 1.3. As a matter of fact, it appears that the probability of misclassification is controlled by two fundamental information quantities: mutual information $I(P_X;Q_{U|X})$ and the cross-entropy loss $\mathcal{L}(Q_{\hat{Y}|U},Q_{U|X})$.

1.3.3 Noisy Lossy Source Coding and the Information Bottleneck

A more subtle variant of the lossy source coding problem is the *noisy lossy source coding problem*, first introduced in [28]. The main difference with respect to the original Shannon's problem relies on that the source Y is not observed directly at the encoder. Instead, a noisy version of Y denoted by X is observed and appropriately compressed. More precisely, we have a memoryless source with single-letter distribution P_Y observed through a noisy channel with single-input transition probability $P_{X|Y}$. From the compressed version of X it is desired to reconstruct, with a predetermined level of fidelity, the realization of the unobserved source Y. The fidelity is measured, similarly to the usual lossy source coding problem, with distortion function $d:\mathcal{Y}\times\mathcal{U}\to\mathbb{R}_{\geq 0}$, where $\mathcal{U}$ is the alphabet in which we generate the reconstructions. Operational information-theoretic definitions for this problem are analogous to Definitions 1.15 and 1.16, and for this reason are omitted. The rate-distortion in this case is given by:

$$\mathcal{R}_{XY,d}(D) = \inf_{\substack{P_{U|X}:\,\mathcal{X}\to\mathcal{P}(\mathcal{U}) \\ \mathbb{E}_{P_{UY}}[d(Y;U)]\leq D}} I\big(P_X;P_{U|X}\big). \tag{1.37}$$

Consider the case of logarithmic distortion $d(y;u) = -\log P_{Y|U}(y|u)$, where

$$P_{Y|U}(y|u) = \frac{\sum_{x\in\mathcal{X}} P_{U|X}(u|x)P_{XY}(x,y)}{\sum_{x\in\mathcal{X}} P_{U|X}(u|x)P_X(x)}. \tag{1.38}$$

The noisy lossy source coding with this choice of distortion function give rise to the celebrated *information bottleneck* [20]. In precise terms:

$$\mathcal{R}_{XY,d}(D) = \inf_{\substack{P_{U|X}\colon \mathcal{X}\to\mathcal{P}(\mathcal{U}) \\ H(P_{Y|U}|P_U)\leq D}} I\big(P_X; P_{U|X}\big). \tag{1.39}$$

Noticing that $H(P_{Y|U}|P_U) = -I\big(P_Y; P_{U|Y}\big) + H(P_Y)$ and defining $\mu \coloneqq H(P_Y) - D$, we can write (1.39) as:

$$\bar{\mathcal{R}}_{XY}(\mu) = \inf_{\substack{P_{U|X}\colon \mathcal{X}\to\mathcal{P}(\mathcal{U}) \\ I(P_Y;P_{U|Y})\geq \mu}} I\big(P_X; P_{U|X}\big). \tag{1.40}$$

Equation (1.40) summarizes the trade-off that exists between the level of compression of the observable source X, using representation U, and the level of information about the *hidden* source Y preserved by this representation. This function is called *rate-relevance* function, where μ is the minimum level of relevance we expect from representation U when the rate used for the compression of X is $\bar{\mathcal{R}}_{XY}(\mu)$. Notice that in the information bottleneck case the distortion $d(y;u)$ depends on the optimal conditional distribution $P^*_{U|X}$ through (1.38). This makes the problem of characterizing $\bar{\mathcal{R}}_{XY}(\mu)$ more difficult than (1.37) in which the distortion function is fixed. In fact, although $\bar{\mathcal{R}}_{XY}(\mu)$ is positive, non-decreasing and convex, the problem in (1.40) is not convex, which leads to the need of more sophisticated tools for its solution. Moreover, from the corresponding operational definition for the lossy source coding problem (analogous to Definitions 1.15 and 1.16) it is clear that the distortion function for sequences Y^n and U^n is applied symbol-by-symbol $\bar{d}(Y^n;U^n) = -\frac{1}{n}\sum_{i=1}\log P_{Y|U}(Y_i|U_i)$, implying a memoryless condition between hidden source realization Y^n and description $U^n = f_n(X^n)$. It is possible to show [29], [30] that if we apply a full logarithmic distortion $\bar{d}(Y^n;U^n) = -\frac{1}{n}\log P_{Y^n|U^n}(Y^n|U^n)$, not necessarily additive as in the previous case, the rate-relevance function in (1.40) remains unchanged, where relevance is measured by the non-additive multi-letter mutual information:

$$\bar{d}(Y^n;U^n) \equiv \frac{1}{n} I\left(P_{Y^n}; P_{f_n(X^n)|Y^n}\right). \tag{1.41}$$

As a simple example in which the rate-relevance function in (1.40) can be calculated in closed form, we can consider the case in which X and Y are jointly Gaussian with zero-mean, variances σ_X^2 and σ_Y^2 and Pearson correlation coefficient given by ρ_{XY}. Using standard information-theoretic arguments [30], it can be shown that the optimal distribution $P_{U|X}$ is also Gaussian with mean X and variance given by:

$$\sigma^2_{U|X} = \sigma_X^2 \frac{2^{-2\mu} - (1-\rho_{XY}^2)}{1-2^{-2\mu}}. \tag{1.42}$$

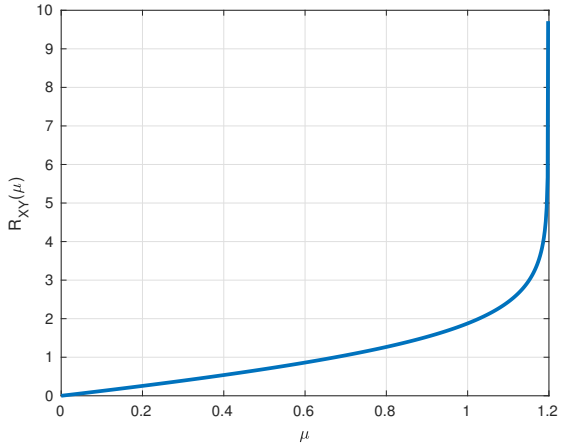

Figure 1.1 $\bar{\mathcal{R}}_{XY}(\mu)$ for $\rho_{XY} = 0.9$.

With this choice for $P_{U|X}$ we easily obtain that $I\big(P_Y; P_{U|Y}\big) = \mu$ and that:

$$\bar{\mathcal{R}}_{XY}(\mu) = \frac{1}{2}\log\left(\frac{\rho_{XY}^2}{2^{-2\mu} - (1-\rho_{XY}^2)}\right), \quad 0 \leq \mu \leq \frac{1}{2}\log\left(\frac{1}{1-\rho_{XY}^2}\right). \tag{1.43}$$

It is interesting to observe that $\bar{\mathcal{R}}_{XY}(\mu)$ depends only on the structure of the sources X and Y through the correlation coefficient ρ_{XY} and not on their variances. It should also be noted that the level of relevance μ is constrained to lie in a bounded interval. This is not surprising because of the Markov chain $U \multimap X \multimap Y$, the maximum value for the relevance level is $I\big(P_X; P_{Y|X}\big)$, which is easily shown to be equal to $\frac{1}{2}\log\left(\frac{1}{1-\rho_{XY}^2}\right)$. The maximum level of relevance can only be achievable as long as the rate $R \to \infty$, that is when the source X is minimally compressed. The trade-off between rate and relevance for this simple example can be appreciated in Fig. 1.1 for $\rho_{XY} = 0.9$.

1.3.4 The Information Bottleneck Method

The noisy lossy source coding with logarithmic loss can be used as a general principle for learning problems leading to the *information bottleneck method*. This method was successfully used in several learning problems with considerably success (see [31]- [32] and references therein). Consider the classification problem introduced in Section 1.2.1 and encoder/decoder pairs $(Q_{U|X}, Q_{\hat{Y}|U})$, as was explained in Section 1.2.2. The information bottleneck method can be introduced through the following optimization problem:

$$\inf_{Q_{U|X}, Q_{\hat{Y}|U}} \Big\{\mathbb{E}_{P_{XY}Q_{U|X}}[-\log Q_{\hat{Y}|U}(Y|U)] + \beta \cdot I\big(P_X; Q_{U|X}\big)\Big\}. \tag{1.44}$$

Expression (1.44) can be interpreted as a cross-entropy loss with a regularization term given by $\beta \cdot I\big(P_X; Q_{U|X}\big)$, where β is positive number. The regularization term can be interpreted as penalization on the complexity of the descriptions generated from the examples X using the encoder $Q_{U|X}$. The smaller the term $I\big(P_X; Q_{U|X}\big)$, the simpler the descriptions U will be. Moreover, as the descriptions U are more simple, they share less information with X and labels Y (because of the Markov chain $U \multimap X \multimap Y$). As the information content in U with respect to Y is naturally decreased, the value of the cross-entropy $\mathbb{E}_{P_{XY}Q_{U|X}}[-\log Q_{\hat{Y}|U}(Y|U)]$ increases. In this way, a trade-off between the cross-entropy loss and the complexity of the descriptions extracted from X is established. It can be though the regularization term given $I\big(P_X; Q_{U|X}\big)$ penalizes very complex descriptions that could provide a low cross-entropy value at the cost of poor generalization and overfitting.

From the result in Lemma 1.5 and the fact that the regularization term $I\big(P_X; Q_{U|X}\big)$ does not depend on the decoder $Q_{\hat{Y}|U}$, problem (1.44) can be written as:

$$\inf_{Q_{U|X}:\mathcal{X}\to\mathcal{P}(\mathcal{U})} \left\{\mathbb{E}_{P_{XY}Q_{U|X}}[-\log Q_{Y|U}(Y|U)] + \beta \cdot I\big(P_X; Q_{U|X}\big)\right\}, \tag{1.45}$$

where the decoder can be written as a function of the encoder as follows:

$$Q_{Y|U}(y|u) = \frac{\sum\limits_{x\in\mathcal{X}} Q_{U|X}(u|x)P_{XY}(x,y)}{\sum\limits_{x\in\mathcal{X}} Q_{U|X}(u|x)P_X(x)}. \tag{1.46}$$

Recognizing that $\mathbb{E}_{P_{XY}Q_{U|X}}[-\log Q_{Y|U}(Y|U)] = H(Q_{Y|U}|Q_U)$, where $Q_U(u) = \sum_{x\in\mathcal{X}} Q_{U|X}(u|x)P_X(x)$, we see that (1.45) is closely related to the information bottleneck and with the rate-relevance function defined in (1.40). In fact, the problem in (1.45) can be equivalently written as:

$$\sup_{Q_{U|X}:\mathcal{X}\to\mathcal{P}(\mathcal{U})} \left\{I\big(P_Y; Q_{U|Y}\big) - \beta \cdot I\big(P_X; Q_{U|X}\big)\right\}, \tag{1.47}$$

with $Q_{U|Y}(u|y) = \sum_{x\in\mathcal{X}} Q_{U|X}(u|x)P_{X|Y}(x|y)$. We can easily see that in (1.47) we are considering the dual problem to (1.40), looking for the supremum of relevance μ subject to a given rate R. The value of β (which can be thought as a typical Lagrange multiplier [33]) can be thought as an hyperparameter which control the trade-off between $I\big(P_Y; Q_{U|Y}\big)$ (relevance) and $I\big(P_X; Q_{U|X}\big)$ (rate). In more precise terms, consider the following set:

$$\begin{aligned}\mathcal{R} \coloneqq \big\{(\mu, R) \in \mathbb{R}^2_{\geq 0} : \exists\, Q_{U|X} : \mathcal{X} \to \mathcal{P}(\mathcal{U}) \ \text{ s.t.} \\ R \geq I\big(P_X; Q_{U|X}\big), \\ \mu \leq I\big(P_Y; Q_{U|Y}\big), \quad U \multimap X \multimap Y\big\}.\end{aligned} \tag{1.48}$$

It is easy to show that this region corresponds to the set of achievable values of relevance and rate (μ, R) for the corresponding noisy lossy source coding problem with logarithmic distortion as was defined in Section 1.3.3. This set is closed and

convex and it is not difficult to show that [34]:

$$\sup_{\substack{Q_{U|X}:\mathcal{X}\to\mathcal{P}(\mathcal{U}) \\ I\left(P_X;Q_{U|X}\right)\leq R}} I\left(P_Y;Q_{U|Y}\right) = \sup\{\mu : (\mu, R) \in \mathcal{R}\}. \tag{1.49}$$

Using convex optimization theory [33], we can easily conclude that (1.47) corresponds to the obtention of the *supporting hyperplane* of region $\mathcal{R}$ with slope β. As any convex and closed set is characterized by all its supporting hyperplanes, varying β and solving (1.47) we are reconstructing the upper boundary of $\mathcal{R}$ which coincides with (1.49). In other words, the hyperparameter β is directly related with the value of R at which we are considering the maximum possible value of redundancy μ, or what it is the same, the value of β controls the complexity of representations of X as was pointed out above.

It only remains to discuss the implementation of a procedure for solving (1.47). Unfortunately, although the set $\mathcal{R}$ characterizing the solutions of (1.47) is convex, it is not true that (1.47) is itself a convex optimization problem. However, the structure of the problem allows for efficient numerical optimization procedures that guarantee convergence to local optimum solutions. These numerical procedures are basically 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 channel and source distributions, respectively [35], [36]. For these reasons, these algorithms can be applied with minor changes to the problem (1.47) as was done in [20].

Clearly, for the solution of (1.47) we need as input the distribution P_{XY}. When only training samples and labels $\mathcal{S}_n \coloneqq \{(x_1, y_1), \ldots, (x_n, y_n)\}$ are available, we use the empirical distribution $\hat{P}_{XY}$ instead of the true distribution P_{XY}.

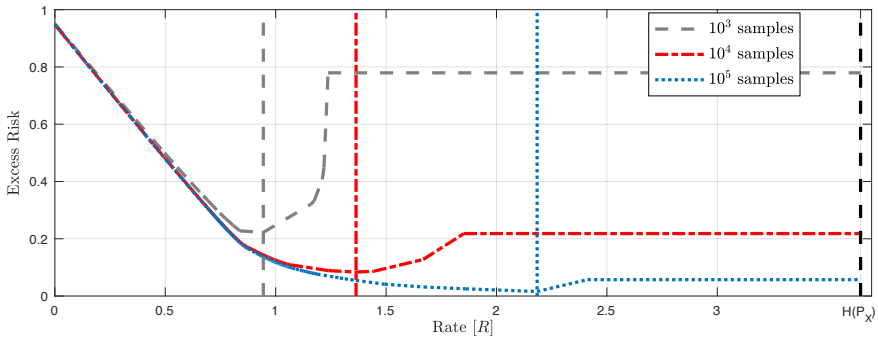

Figure 1.2 Excess risk (1.50) as a function of rate R being the mutual information between the representation U and the corresponding input X.

In Fig. 1.2, we plot what we call the excess-risk (as presented in Definition 1.14) rewritten below as:

$$\text{Excess-risk} \coloneqq H(Q^{*,\beta}_{Y|U}|Q^{*,\beta}_{U}) - H(P_{Y|X}|P_X), \tag{1.50}$$

where $Q^{*,\beta}_{Y|U}$, $Q^{*,\beta}_{U}$ are computed by using the optimal solution $Q^{*,\beta}_{U|X}$ in (1.47) and the empirical distribution $\hat{P}_{XY}$. As β defines unequivocally the value of $I\big(P_X; Q^{*,\beta}_{U|X}\big)$, which is basically the rate or complexity associated with the chosen encoder, we choose the horizontal axis to be labeled by rate R. Experiments were performed by using synthetic data with alphabets $|\mathcal{X}| = 128$ and $|\mathcal{Y}| = 4$. Excess risk curve as a function of the rate constraint for different sizes of training samples is plotted. With dash vertical coloured 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: $H(P_X)$ (vertical line 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. It is not difficult to check that $R_{\text{lim}} = H(\hat{P}_X)$. Furthermore, for every size of training samples, there is an optimal value of R_{opt} which provides the lowest excess-risk in (1.50). 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. This fact was expected since when the amount of training data increases the empirical distribution approaches the data-generating distribution.

In the next section, we provide a formal mathematical proof of the explicit relation between the generalization gap and the rate constraint, which explains the heuristics observations presented in Fig. 1.2.

1.4 The Interplay Between Information and Generalization

In the following we will denote $\mathcal{L}(Q_{U|X}) \equiv \mathcal{L}(Q_{U|X}, \hat{Q}_{Y|U})$ and $\mathcal{L}_{\text{emp}}(Q_{U|X}) \equiv \mathcal{L}_{\text{emp}}(Q_{U|X}, \hat{Q}_{Y|U})$. We will study informational bounds on the generalization gap (1.18). More precisely, the goal is to find the learning rate $\epsilon_n(Q, \mathcal{S}_n, \gamma_n)$ such that

$$\mathbb{P}\left(\mathcal{E}_{\text{gap}}(Q_{U|X}, \mathcal{S}_n) > \epsilon_n(Q_{U|X}, \mathcal{S}_n, \gamma_n)\right) \leq \gamma_n, \tag{1.51}$$

for a given $Q_{U|X} \in \mathcal{F}$ and some $\gamma_n \to 0$ as $n \to \infty$. We will further comment implications for practical algorithms minimizing the surrogate of the risk function:

$$\mathcal{L}(Q_{U|X}) \leq \mathcal{L}_{\text{emp}}(Q_{U|X}) + \mathcal{E}_{\text{gap}}(Q_{U|X}), \tag{1.52}$$

which depends on the empirical risk and the so-called generalization gap. Expression (1.52) states that a suitable selection of the encoder can be obtained by minimizing the empirical risk and the generalization gap simultaneously, that is:

$$\mathcal{L}_{\text{emp}}\big(Q_{U|X}\big) + \lambda \cdot \epsilon_n(Q_{U|X}, \mathcal{S}_n, \gamma_n), \tag{1.53}$$

for some suitable multiplier $\lambda \geq 0$. It is reasonable to expect that the optimal encoder achieving the minimal risk in (1.10) does not belong to $\mathcal{F}$, so we may want to enlarge the model classes as much as possible. However, as usual, we expect a sensitive tradeoff between these two fundamental quantities.

1.4.1 Bounds on the Generalization Gap

We first present the main technical result in Theorem 1.20, that is a sample-dependent bound on the generalization gap (1.18) with probability of at least $1-\delta$, as a function of a selected randomized encoder $Q_{U|X}$ and the data probability distribution P_{XY}. In particular, we will show that the mutual information between raw data and its representation controls the learning rate with an order $\mathcal{O}\left(\frac{\log(n)}{\sqrt{n}}\right)$, which leads to an informational PAC style generalization error bound. From this perspective, we discuss implications for model selection, variational auto-encoders and the information bottleneck (IB) method.

THEOREM 1.20 (Informational bound) *Let $\mathcal{F}$ be a class of encoders. Then, for every P_{XY} and every $\delta \in (0,1)$, with probability at least $1-\delta$ over the choice of $\mathcal{S}_n \sim P_{XY}^n$ the following inequality holds $\forall\, Q_{U|X} \in \mathcal{F}$:*

$$\mathcal{E}_{gap}(Q_{U|X}, \mathcal{S}_n) \leq A_\delta \sqrt{I(\hat{P}_X; Q_{U|X})} \cdot \frac{\log(n)}{\sqrt{n}} + \frac{C_\delta}{\sqrt{n}} + \mathcal{O}\left(\frac{\log(n)}{n}\right), \tag{1.54}$$

where $(A_\delta, B_\delta, C_\delta)$ are universal constants:

$$A_\delta \coloneqq \frac{\sqrt{2}B_\delta}{P_X(x_{\min})}\big(1 + 1/\sqrt{|\mathcal{X}|}\big), \quad B_\delta \coloneqq 2 + \sqrt{\log\left(\frac{|\mathcal{Y}|+3}{\delta}\right)}, \tag{1.55}$$

$$C_\delta \coloneqq 2|\mathcal{U}|e^{-1} + B_\delta\sqrt{|\mathcal{Y}|}\log\frac{|\mathcal{U}|}{P_Y(y_{\min})}. \tag{1.56}$$

The importance of this result is that the main quantity involves the empirical mutual information between data X and its randomized representation $U(X)$. This can be understood as a "measure of information complexity" scaling with rate $n^{-1/2}\log(n)$. The remaining issue is merely how to interpret this information-theoretic bound and its implication in the learning problem.

By combining Theorem 1.20 with inequality (1.52) we obtain the following corollary.

COROLLARY (PAC style generalization error bound) *Let $\mathcal{F}$ be the class of randomized encoders. Then, for every P_{XY} and every $\delta \in (0,1)$, with probability at least $1-\delta$ over the choice of $\mathcal{S}_n \sim P_{XY}^n$ the following inequality holds:*

$$\mathcal{L}(\hat{Q}_{Y|U}, Q_{U|X}) \leq H(\hat{Q}_{Y|U}|\hat{Q}_U) + A_\delta\sqrt{I(\hat{P}_X; Q_{U|X})}\frac{\log(n)}{\sqrt{n}} + \frac{C_\delta}{\sqrt{n}} + \mathcal{O}\left(\frac{\log(n)}{n}\right). \tag{1.57}$$

An interesting connection between the empirical risk minimization of the cross-entropy loss and the information bottleneck method presented in the previous section arises which motivates formally the following algorithm [15, 20, 37].

DEFINITION 1.21 (Information bottleneck algorithm) A representation learning algorithm inspired by the information bottleneck principle [20] consists in finding an encoder $Q_{U|X} \in \mathcal{F}$ that minimizes over the random choice $\mathcal{S}_n \sim P_{XY}^n$ the functional:

$$\mathcal{L}_{\mathrm{IB}}^{(\lambda)}(Q_{U|X}) \coloneqq H(\hat{Q}_{Y|U}|\hat{Q}_U) + \lambda \cdot I(\hat{P}_X; Q_{U|X}), \tag{1.58}$$

for a suitable multiplier $\lambda > 0$, where $\hat{Q}_{Y|U}$ is given by (1.16) and $\hat{Q}_U$ is its denominator.

This algorithm optimizes a tradeoff between $H(\hat{Q}_{Y|U}|\hat{Q}_U)$ and the information-based regularization term $I(\hat{P}_X; Q_{U|X})$. Interestingly, the resulting regularized empirical risk suggested by (1.57) can be seen as an optimization of the IB method from the empirical distribution (1.58) but based on the square-root of the mutual information in expression (1.58). Additionally, we observe that by selecting an arbitrary $\tilde{Q}_U \in \mathcal{P}(\mathcal{U})$ in (1.58) and using the information-radius identity [24], the next inequality holds:

$$\mathcal{L}_{\mathrm{IB}}^{(\lambda)}(Q_{U|X}) \leq H(\hat{Q}_{Y|U}|\hat{Q}_U) + \lambda \cdot D\big(Q_{U|X}(\cdot|X)\|\tilde{Q}_U|\hat{P}_X\big) \tag{1.59}$$

$$\equiv \mathcal{L}_{\mathrm{VA}}^{(\lambda)}(Q_{U|X}, \tilde{Q}_U). \tag{1.60}$$

The new surrogate function (1.60), denoted by $\mathcal{L}_{\mathrm{VA}}^{(\lambda)}(Q_{U|X})$, shares a lot of in common with a slightly more general form of the variational auto-encoders (VAEs) [16] and the recently introduced Information Dropout (ID) [15,37], where the latent space is regularized using a prior $\tilde{Q}_U$. Therefore, the information-theoretic bound in Theorem 1.20 shows that the algorithm in Definition 1.21 as well as VAEs and ID are slightly different but related information-theoretic ways to control the generalization gap. In all of them the mutual information $I(\hat{P}_X; Q_{U|X})$ (or its upper bound given by $D\big(Q_{U|X}(\cdot|X)\|\tilde{Q}_U|\hat{P}_X\big)$) plays the fundamental role, although the specific way in which this term control the generalization gap could be different for each case.

1.4.2 Information Complexity of Representations

We could think of the most significative term in the upper bound (1.54) as an information complexity cost of data representations, which depends only on the data samples and on the selected randomized encoder from the restricted model. Suppose we are given with a set of different model classes for the randomized encoders $k = [1 : K]$:

$$\mathcal{F}_E^{(k)} \coloneqq \Big\{ Q_{U|X} \equiv \mathbb{E}_{P_Z}\big[\mathbb{1}[u = f_{\boldsymbol{\theta}}(x, Z)]\big] \,:\, \boldsymbol{\theta} = (\theta_1, \ldots, \theta_k) \in \Theta_k,\, P_Z \in \mathcal{P}_k(\mathcal{Z}) \Big\},$$

where there are two kinds of parameters: a structure parameter k and real-value parameters θ, whose parameters depend on the structure, e.g., Θ_k may account for different number of layers or non-linearities while $\mathcal{P}_k(\mathcal{Z})$ indicates different kind of noise distributions. Theorem 1.20 motivates the following model selection principle for learning compact representations:

Find a parameter k and real-value parameters $\boldsymbol{\theta}$ for the observed data $\mathcal{S}_n$ with which the corresponding data representation can be encoded with the shortest code length:

$$\inf_{\boldsymbol{\theta}\in\Theta_k\,,\,k=[1:K]} \left[\mathcal{L}_{emp}\left(Q_{U|X}^{(\boldsymbol{\theta},k)}, \mathcal{S}_n\right) + \lambda \cdot \sqrt{I\left(\hat{P}_X; Q_{U|X}^{(\boldsymbol{\theta},k)}\right)} \right], \tag{1.61}$$

where the mutual information penalty term indicates the minimum of the expected redundancy between the minimum code-length[5] *(measured in bits) $-\log Q_{U|X}^{(\boldsymbol{\theta},k)}(\cdot|x)$ to encode representations under a known data source and the best code-length $-\log Q_U(\cdot)$ chosen to encode the data representations without knowing the input samples:*

$$I\left(\hat{P}_X; Q_{U|X}^{(\boldsymbol{\theta},k)}\right) = \min_{Q_U\in\mathcal{P}(\mathcal{U})} \mathbb{E}_{\hat{P}_X}\mathbb{E}_{Q_{U|X}^{(\boldsymbol{\theta},k)}}\left[-\log Q_U(U) + \log Q_{U|X}^{(\boldsymbol{\theta},k)}(U|X)\right]. \tag{1.62}$$

This information principle combines the empirical cross-entropy risk (1.14) with the "information complexity" of the selected encoder (1.62) as being a regularization that acts as a sample-dependent penalty against *overfitting*. One may view (1.62) as a possible means of comparing the appropriateness of distinct representation models (e.g., number of layers or amount of noise), after a parametric choice has been selected.

The coding interpretation of the penalty term in (1.61) is that the length of the description of the representations themselves can be quantified in the same units as the code length in data compression, namely, bits. In other words, for each data sample x, a randomized encoder can induce different types of representations $U(x)$ with expected information length given by $H\left(Q_{U|X}(\cdot|x)\right)$. When this representation has to be encoded without knowing $Q_{U|X}$ since x is not given to us (e.g. a communication problem where the sender wishes to communicate the representations only), the required average length of an encoding distribution Q_U results in $\mathbb{E}_{Q_{U|X}}\left[-\log Q_U(U)\right]$. In this sense, expression (1.61) suggests to select encoders that allow us to then encode representations efficiently. Interestingly, this is closely related to the celebrated minimum description length (MDL) method for density estimation [38, 39]. However, the fundamental differences between these principles is that the information complexity (1.62) follows from the generalization gap and measures the amount of information conveyed by the representations relative to an encoder model, as opposed to the model parameters of the encoder itself.

The information-theoretic significance of (1.62) goes beyond a simply regularization term since it leads to us to introduce the fundamental notion of *encoder capacity*. This key idea of encoder capacity is made possible thanks to Theorem 1.20 that connects mathematically the generalization gap to the information

[5] As is well known in information theory, the shortest expected code length achievable by a uniquely decodable code under a known data source [24].

complexity which is intimately related to the number of distinguishable samples from the representations. Notice that the information complexity can be upper bounded as:

$$I\left(\hat{P}_X; Q_{U|X}\right) = \frac{1}{n}\sum_{i=1}^{n} D\left(Q_{U|X}(\cdot|x_i)\middle\|\frac{1}{n}\sum_{j=1}^{n} Q_{U|X}(\cdot|x_j)\right) \tag{1.63}$$

$$\leq \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} D\left(Q_{U|X}(\cdot|x_i)\middle\|Q_{U|X}(\cdot|x_j)\right), \tag{1.64}$$

where $\{x_i\}_{i=1}^n$ are the training examples from the dataset $\mathcal{S}_n$ and the last inequality follows from the convexity of the relative entropy. This bound is measuring the average degree of closeness between the corresponding representations for the different sample inputs. When two distributions, $Q_{U|X}(\cdot|x_i)$ and $Q_{U|X}(\cdot|x_j)$ are very close to each other, i.e., $Q_{U|X}$ assigns high likelihood to similar representations corresponding to different inputs $x_i \neq x_j$, they do not contribute so much to the complexity of the overall representations. In other words, the more sample inputs an encoder can differentiate, the more patterns it can fit well so the larger the mutual information will be and thus the risk of overfitting. This observation suggests that the complexity of a representation model with respect to a sample dataset can be related to the number of data samples that essentially yields different (distinguishable) representations. Inspired by the concept of *stochastic complexity* [39], we introduce below the notion of *encoder capacity* to measure the complexity of a representation model:

DEFINITION 1.22 (Capacity of randomized encoders) The *encoder capacity* $\mathcal{C}_e$ of a randomized encoder $Q_{U|X}$ with respect to a sample set $\mathcal{A} \subseteq \mathcal{X}$ is defined as:

$$\mathcal{C}_e(\mathcal{A}, Q_{U|X}) \coloneqq \max_{\psi\,:\,\mathcal{U}\to\mathcal{A}} \log\left(\sum_{u\in\mathcal{U}} Q_{U|X}\big(u|\psi(u)\big)\right) = \log|\mathcal{A}| - \log\left(\frac{1}{1-\varepsilon}\right), \tag{1.65}$$

$$\varepsilon \coloneqq \min_{\psi\,:\,\mathcal{U}\to\mathcal{A}} \frac{1}{|\mathcal{A}|}\sum_{x\in\mathcal{A}}\sum_{u\in\mathcal{U}} Q_{U|X}\left(u|x\right)\mathbb{1}\left[\psi(u)\neq x\right] \leq 1 - \frac{1}{|\mathcal{A}|}. \tag{1.66}$$

The argument of the logarithm in the second term of (1.65) represents the probability to distinguish samples from their representations $1-\varepsilon$, i.e., the average probability that estimated samples via the maximum-likelihood estimator $\psi(\cdot)$ from $Q_{U|X}$ be equal to the true samples. Therefore, the encoder capacity is the logarithm of the number of total samples minus a term that depends on the misclassification probability of the input samples from their representations. When ε is small, then $\mathcal{C}_e(\mathcal{A}, Q_{U|X}) \approx \log|\mathcal{A}| - \varepsilon$ and thus, all samples are perfectly distinguishable. The following proposition gives simple bounds[6] on the

[6] Notice that it is possible to provide better bounds on ε by relaying on the results in [40]. However, we preferred simplicity to "tightness" since the purpose of Proposition 1.23 is to link the encoder capacity and the information complexity.

encoder capacity from the information complexity (1.62) which as we already know has a close relation with the generalization gap:

PROPOSITION 1.23 *Let $Q_{U|X}$ be an encoder distribution and $\hat{P}_X$ be an empirical distribution with support $\mathcal{A}_n \equiv supp(\hat{P}_X)$. Then, the information complexity and the encoder capacity satisfy:*

$$\mathcal{C}_e\big(\mathcal{A}_n, Q_{U|X}\big) = \log|\mathcal{A}_n| - \log\left(\frac{1}{1-\varepsilon}\right) \tag{1.67}$$

$$g^{-1}\left(\log|\mathcal{A}_n| - I\big(\hat{P}_X; Q_{U|X}\big)\right) \leq \varepsilon \leq \frac{1}{2}\left(\log|\mathcal{A}_n| - I\big(\hat{P}_X; Q_{U|X}\big)\right), \tag{1.68}$$

where ε is defined by (1.66) with respect to $\mathcal{A}_n$ and, for $0 \leq t \leq 1$,

$$g(t) \coloneqq t \cdot \log\left(|\mathcal{A}_n| - 1\right) + h(t) \tag{1.69}$$

with $h(t) \coloneqq -t\log(t) - (1-t)\log(1-t)$ and $0\log 0 \coloneqq 0$. Furthermore,

$$I\big(\hat{P}_X; Q_{U|X}\big) \leq \mathcal{C}_e. \tag{1.70}$$

Proof We begin with the lower bound (1.70). Consider the inequalities:

$$I\big(\hat{P}_X; Q_{U|X}\big) = \min_{Q_U \in \mathcal{P}(\mathcal{U})} D\big(Q_{U|X}\big\|Q_U\big|\hat{P}_X\big) \tag{1.71}$$

$$\leq \min_{Q_U \in \mathcal{P}(\mathcal{U})} \mathbb{E}_{\hat{P}_X}\mathbb{E}_{Q_{U|X}}\left[\max_{x \in \mathcal{A}_n} \log \frac{Q_{U|X}(U|x)}{Q_U(U)}\right] \tag{1.72}$$

$$\leq \min_{Q_U \in \mathcal{P}(\mathcal{U})} \max_{u \in \mathcal{U}} \log \frac{Q_{U|X}\big(u|\psi^\star(u)\big)}{Q_U(u)} \tag{1.73}$$

$$= \log\left(\sum_{u \in \mathcal{U}} Q_{U|X}\big(u|\psi^\star(u)\big)\right) = \mathcal{C}_e\big(Q_{U|X}, \mathcal{A}_n\big), \tag{1.74}$$

where inequality (1.73) follows by letting $\psi^\star$ to be the mapping maximizing $\mathcal{C}_e\big(Q_{U|X}, \mathcal{A}_n\big)$ and (1.74) follows by noticing that (1.73) is the smallest worst-case regret, known as the *minimax regret*, and thus by choosing Q_U to be the normalized maximum-likelihood distribution on the restricted set $\mathcal{A}_n$ the claim is a consequence of remarkable result of Shtarkov [41].

It remains to show the bounds in (1.68). In order to show the lower bound, we can simply apply Fano's lemma [42, Lemma 2.10] from which we can bound from below the error probability (1.66) based on $\mathcal{A}_n$. As for the upper bound,

$$\log|\mathcal{A}_n| - I\big(\hat{P}_X; Q_{U|X}\big) \geq H\big(\hat{P}_X\big) - I\big(\hat{P}_X; Q_{U|X}\big) \tag{1.75}$$

$$= \sum_{u \in \mathcal{U}} \hat{Q}_U(u) H\big(\hat{Q}_{X|U}(\cdot|u)\big) \tag{1.76}$$

$$\geq 2\sum_{u \in \mathcal{U}} \hat{Q}_U(u)\left(1 - \max_{x' \in \mathcal{X}} \hat{Q}_{X|U}(x'|u)\right) \tag{1.77}$$

$$= 2\varepsilon, \tag{1.78}$$

where (1.75) follows from the assumption $\mathcal{A}_n = \mathrm{supp}\big(\hat{P}_X\big)$ and the fact that the

entropy is maximal over the uniform distribution; (1.77) follows by using [43, eq. (7)] and (1.78) by the definition of ε in (1.66). This concludes the proof. □

Remark 1.1 In Proposition 1.23, the function $g^{-1}(t) := 0$ for $t < 0$ and, for $0 < t < \log|\mathcal{A}_n|$, $g^{-1}(t)$ is a solution of the equation $g(\varepsilon) = t$ with respect to $\varepsilon \in \left[0, 1 - 1/|\mathcal{A}_n|\right]$; this solution exists since the function g is continuous and increasing on $\left[0, 1 - 1/|\mathcal{A}_n|\right]$ and $g(0) = 0$, $g\big(1 - 1/|\mathcal{A}_n|\big) = \log|\mathcal{A}_n|$.

Remark 1.2 (Generalization requires learning invariant representations) An important consequence of the lower bound in (1.68) in Proposition 1.23 is that by limiting the information complexity, i.e., by controlling the generalization gap according to the criterion (1.61), we bound from below the error probability of distinguishing input samples from their representations. In other words, from expression (1.67) and Theorem 1.20 we can conclude that encoders inducing a large misclassification probability on input samples from their representations, i.e., different inputs must share similar representations, are expected to achieve better generalization. Specifically, this also implies formally that we only need to enforce invariant representations to control the encoder capacity (e.g., injecting noise during training), from which the generalization is upper bounded naturally thanks to Theorem 1.20 and the connection with the information complexity. However, there is a sensitive tradeoff between the amount of noise (enforcing both invariance and generalization) and the minimization of the cross-entropy loss. Additionally, it is not difficult to show from the data-processing inequality that stacking noisy encoder layers reinforce increasingly invariant representations since distinguishing inputs from their representations becomes harder –or equivalently the encoder capacity decreases– as the network is deeper.

1.4.3 Sketch of the Proofs

We begin by observing that the generalization gap can be easily bounded as:

$$\begin{aligned}\mathcal{E}_{\text{gap}}(Q_{U|X}, \mathcal{S}_n) \leq \tilde{\mathcal{E}}_{\text{gap}}(Q_{U|X}, \mathcal{S}_n) \\ + \Big| \sum_{(u,y)\in\mathcal{U}\times\mathcal{Y}} \Big[Q_{YU}(y,u) - \hat{Q}_{YU}(y,u)\Big] \log\left(\frac{Q_{Y|U}(y|u)}{\hat{Q}_{Y|U}(y|u)}\right)\Big|,\end{aligned} \tag{1.79}$$

where we define:

$$\begin{aligned}\tilde{\mathcal{E}}_{\text{gap}}(Q_{U|X}, \mathcal{S}_n) = \Big|\mathbb{E}_{P_{XY}}\left[-\sum_{u\in\mathcal{U}} Q_{U|X}(u|x)\log Q_{Y|U}(Y|U)\right] \\ - \mathbb{E}_{\hat{P}_{XY}}\left[-\sum_{u\in\mathcal{U}} Q_{U|X}(u|x)\log Q_{Y|U}(Y|U)\right]\Big|.\end{aligned} \tag{1.80}$$

That is, $\tilde{\mathcal{E}}_{\text{gap}}(Q_{U|X}, \mathcal{S}_n)$ is the gap corresponding to the optimal decoder selecting, which depends on the true P_{XY}, according to Lemma 1.5. It is not difficult

to show that

$$\tilde{\mathcal{E}}_{\text{gap}}(Q_{U|X}, \mathcal{S}_n) \leq \left| H(Q_{Y|U}|Q_U) - H(\hat{Q}_{Y|U}|\hat{Q}_U) \right| + \mathbb{E}_{\hat{Q}_U}\left[D\big(\hat{Q}_{Y|U} \| Q_{Y|U}\big) \right], \tag{1.81}$$

where the second term can be bounded as $\mathbb{E}_{\hat{Q}_U}\left[D\big(\hat{Q}_{Y|U} \| Q_{Y|U}\big) \right] \leq D\big(\hat{P}_{XY} \| P_{XY}\big)$. The first term of (1.81) is bounded as:

$$\begin{aligned}\left| H(Q_{Y|U}|Q_U) - H(\hat{Q}_{Y|U}|\hat{Q}_U) \right| \leq \left| H(Q_U) - H(\hat{Q}_U) \right| + \left| H(P_Y) - H(\hat{P}_Y) \right| \\ + \left| H(Q_{U|Y}|P_Y) - H(\hat{Q}_{U|Y}|\hat{P}_Y) \right|. \end{aligned} \tag{1.82}$$

To obtain an upper bound, we use the following bounds [18]:

$$\left| H(Q_U) - H(\hat{Q}_U) \right| \leq \sum_{u \in \mathcal{U}} \phi\left(\|\mathbf{p}_X - \hat{\mathbf{p}}_X\|_2 \cdot \sqrt{\mathbb{V}\big(\{Q_{U|X}(u|x)\}_{x \in \mathcal{X}}\big)} \right), \tag{1.83}$$

$$\begin{aligned}&\left| H(Q_{U|Y}|P_Y) - H(\hat{Q}_{U|Y}|\hat{P}_Y) \right| \leq \|\mathbf{p}_Y - \hat{\mathbf{p}}_Y\|_2 \sqrt{|\mathcal{Y}|} \log |\mathcal{U}| \\ &+ \mathbb{E}_{P_Y}\Big[\sum_{u \in \mathcal{U}} \phi\left(\big\|\mathbf{p}_{X|Y}(\cdot|Y) - \hat{\mathbf{p}}_{X|Y}(\cdot|Y)\big\|_2 \cdot \sqrt{\mathbb{V}\big(\{\mathbf{q}_{U|X}(u|x)\}_{x \in \mathcal{X}}\big)} \right) \Big] \end{aligned} \tag{1.84}$$

where

$$\phi(x) = \begin{cases} 0 & x \leq 0 \\ -x \log(x) & 0 < x < e^{-1} \\ e^{-1} & x \geq e^{-1} \end{cases} \tag{1.85}$$

and $\mathbb{V}(\mathbf{a}) = \|\mathbf{a} - \bar{a}\mathbb{1}_d\|_2^2$ with $\mathbf{a} \in \mathbb{R}^d$, $d \in \mathbb{N}_+$, $\bar{a} = \frac{1}{d}\sum_{i=1}^d a_i$, and $\mathbb{1}_d$ is the vector of ones of length d.

It is clear that $P_Y \mapsto H(P_Y)$ is a differentiable function and thus, we can apply a first order Taylor expansion to obtain:

$$H(P_Y) - H(\hat{P}_Y) = \left\langle \frac{\partial H(P_Y)}{\partial \mathbf{p}_Y}, \mathbf{p}_Y - \hat{\mathbf{p}}_Y \right\rangle + o\left(\|\mathbf{p}_Y - \hat{\mathbf{p}}_Y\|_2 \right), \tag{1.86}$$

where $\frac{\partial H(P_Y)}{\partial P_Y(y)} = -\log P_Y(y) - \log(e)$ for each $y \in \mathcal{Y}$. Then using Cauchy-Schwartz inequality we have:

$$\left| H(P_Y) - H(\hat{P}_Y) \right| \leq \sqrt{\mathbb{V}\big(\{\log \mathbf{p}_Y(y)\}_{y \in \mathcal{Y}}\big)}\, \|\mathbf{p}_Y - \hat{\mathbf{p}}_Y\|_2 + o\left(\|\mathbf{p}_Y - \hat{\mathbf{p}}_Y\|_2 \right). \tag{1.87}$$

The McDiarmid's concentration inequality and [24, Theorem 12.2.1] allow us to bound with an arbitrary probability close to one the terms: $D\big(\hat{P}_{XY} \| P_{XY}\big)$, $\|\mathbf{p}_X - \hat{\mathbf{p}}_X\|_2$, $\|\mathbf{p}_Y - \hat{\mathbf{p}}_Y\|_2$ and $\|\mathbf{p}_{X|Y}(\cdot|y) - \hat{\mathbf{p}}_{X|Y}(\cdot|y)\|_2$, $\forall\, y \in \mathcal{Y}$ simultaneously. To make sure the bounds hold simoultaneously over these $\mathcal{Y} + 3$ quantities, we replace δ with $\delta/(|\mathcal{Y}|+3)$ in each concentration inequality. Then, with probability

at least $1-\delta$ the following bounds hold:

$$\max\left\{\left\|\mathbf{p}_X-\hat{\mathbf{p}}_X\right\|_2,\left\|\mathbf{p}_{X|Y}(\cdot|y)-\hat{\mathbf{p}}_{X|Y}(\cdot|y)\right\|_2,\left\|\mathbf{p}_Y-\hat{\mathbf{p}}_Y\right\|_2\right\}\leq\frac{B_\delta}{\sqrt{n}},\quad(1.88)$$

$$D\big(\hat{P}_{XY}\|P_{XY}\big)\leq|\mathcal{X}||\mathcal{Y}|\frac{\log(n+1)}{n}+\frac{1}{n}\log\left(\frac{|\mathcal{Y}|+3}{\delta}\right).\quad(1.89)$$

Then, with probability at least $1-\delta$ we have:

$$\begin{aligned}\tilde{\mathcal{E}}_{\text{gap}}(Q_{U|X},\mathcal{S}_n)&\leq 2\sum_{u\in\mathcal{U}}\phi\left(\frac{B_\delta}{\sqrt{n}}\sqrt{\mathbb{V}(\{\mathbf{q}_{U|X}(u|x)\}_{x\in\mathcal{X}})}\right)+\frac{B_\delta}{\sqrt{n}}\sqrt{|\mathcal{Y}|}\log|\mathcal{U}|\\&\quad+\sqrt{\mathbb{V}\big(\{\log\mathbf{p}_Y(y)\}_{y\in\mathcal{Y}}\big)}\frac{B_\delta}{\sqrt{n}}+\mathcal{O}\left(\frac{\log(n)}{n}\right)\quad(1.90)\\&\leq\frac{\log(n)}{\sqrt{n}}B_\delta\sum_{u\in\mathcal{U}}\sqrt{\mathbb{V}(\{\mathbf{q}_{U|X}(u|x)\}_{x\in\mathcal{X}})}+\frac{B_\delta\sqrt{|\mathcal{Y}|}\log|\mathcal{U}|}{\sqrt{n}}\\&\quad+\frac{2|\mathcal{U}|e^{-1}}{\sqrt{n}}+\sqrt{\mathbb{V}\big(\{\log\mathbf{p}_Y(y)\}_{y\in\mathcal{Y}}\big)}\frac{B_\delta}{\sqrt{n}}+\mathcal{O}\left(\frac{\log(n)}{n}\right),\quad(1.91)\end{aligned}$$

where we use $n\geq a^2e^2$ and $\phi\left(\frac{a}{\sqrt{n}}\right)\leq\frac{a}{2}\frac{\log(n)}{\sqrt{n}}+\frac{e^{-1}}{\sqrt{n}}$. By combining this result with the next inequality [18]:

$$\sum_{u\in\mathcal{U}}\sqrt{\mathbb{V}\big(\{\mathbf{q}_{U|X}(u|x)\}_{x\in\mathcal{X}}\big)}\leq\frac{\sqrt{2}}{p_X(x_{\min})}\left(1+\sqrt{\frac{1}{|\mathcal{X}|}}\right)\sqrt{I(P_X;Q_{U|X})},\quad(1.92)$$

we relate to the mutual information. Finally, using Taylor arguments as above, we can easily write:

$$\left|\sqrt{I\left(P_X;Q_{U|X}\right)}-\sqrt{I\left(\hat{P}_X;Q_{U|X}\right)}\right|\equiv\mathcal{O}(\|\mathbf{p}_X-\hat{\mathbf{p}}_X\|_2)\leq\mathcal{O}(n^{-1/2})\quad(1.93)$$

with probability $1-\delta$. It only remains to analyze the second term in the RHS of (1.79). Using standard manipulations, we can easily show this term can be equivalently written as:

$$\Big|\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}}\Big[P_{XY}(x,y)-\hat{P}_{XY}(x,y)\Big]\sum_{u\in\mathcal{U}}Q_{U|X}(u|x)\log\left(\frac{Q_{Y|U}(y|u)}{\hat{Q}_{Y|U}(y|u)}\right)\Big|.\quad(1.94)$$

It is not difficult to see that given $Q_{U|X}$, $P_{XY}\mapsto\log\big(Q_{Y|U}(y|u)\big)$ is a differentiable function and thus, we can apply a first order Taylor expansion to obtain:

$$\begin{aligned}\sum_{u\in\mathcal{U}}Q_{U|X}(u|x)\log\left(\frac{Q_{Y|U}(y|u)}{\hat{Q}_{Y|U}(y|u)}\right)&=-\sum_{u\in\mathcal{U}}Q_{U|X}(u|x)\left\langle\frac{\partial\log Q_{Y|U}(y|u)}{\partial\mathbf{p}_{XY}},\mathbf{p}_{XY}-\hat{\mathbf{p}}_{XY}\right\rangle\\&\quad+o\left(\|\mathbf{p}_{XY}-\hat{\mathbf{p}}_{XY}\|_2\right),\quad(1.95)\end{aligned}$$

and

$$\frac{\partial\log Q_{Y|U}(y|u)}{\partial P_{XY}(x',y')}=\frac{Q_{U|X}(u|x')\left[\mathbb{1}\{y'=y\}-Q_{Y|U}(y|u)\right]}{Q_{UY}(u,y)}.\quad(1.96)$$

With the assumption that for every encoder $Q_{U|X}(u|x)$ in the family $\mathcal{F}$ satisfying that $Q_{U|X}(u|x) > \alpha$ for every $(u, x) \in \mathcal{U} \times \mathcal{X}$ with $\alpha > 0$, we obtain that:

$$\Big|\frac{\partial \log Q_{Y|U}(y|u)}{\partial P_{XY}(x', y')}\Big| \leq \frac{2}{\alpha},\ \forall (x, x', y', u) \in \mathcal{X} \times \mathcal{X} \times \mathcal{Y} \times \mathcal{U}. \tag{1.97}$$

From simple algebraic manipulations, we can bound the term in (1.94) as

$$\Big| \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \Big[P_{XY}(x,y) - \hat{P}_{XY}(x,y)\Big] \sum_{u\in\mathcal{U}} Q_{U|X}(u|x) \log\left(\frac{Q_{Y|U}(y|u)}{\hat{Q}_{Y|U}(y|u)}\right)\Big| \tag{1.98}$$

$$\leq \frac{2}{\alpha}\left(\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} |P_{XY}(x,y) - \hat{P}_{XY}(x,y)|\right)^2. \tag{1.99}$$

Again, using McDiarmid's concentration inequality, it can be shown that with probability close to one this term is $\mathcal{O}(1/n)$ which can be neglected compared to the other terms previously calculated. This concludes the proof of the theorem.

1.5 Summary and Outlook

We discussed how generalization in representation learning based on the cross-entropy loss is related to the notion of information complexity, and how this connection is employed to view learning in terms of the information bottleneck principle. The resulting information complexity penalty is a sample-dependent bound on the generalization gap that crucially depends on the mutual information between the inputs and the randomized (representations) outputs of the selected encoder, revealing an interesting connection between the generalization capabilities of representation models and the information carried by the representations. Furthermore, we have shown that information complexity is closely related to the so-called encoder capacity revealing the well-known fact that enforcing invariance in the representations is a critical aspect to control the generalization gap. Among other things the results of this chapter present a new viewpoint on the foundations of representation learning, showing the usefulness of information-theoretic concepts and tools in the comprehension of fundamental learning problems. This survey provided a summary of some useful links between information theory and representation learning from which we expect to see advances in years to come.

In the present analysis, the number of samples is the most useful resource for the reduction of the generalization gap. Nevertheless, we have not considered other important ingredients of the problem related to the computational complexity aspect of learning representation models. One of them is the particular optimization problem that has to be solved in order to find an appropriate encoder. It is well-known that the specific "landscape" of the cost function (as function of the parameters of the encoders family) to be optimized and the particular optimization algorithm used (e.g. stochastic gradient descent algortihms)

could have some major effects that may not be improved by increasing the number of samples. Additional constraints imposed by real-world applications such that computations must be performed with a limited time budget could also be relevant from more practical perspective. Evidently, it is pretty clear that many challenges still remain in this exciting research area.

References

[1]CAM, *Frontiers in Massive Data Analysis*. National Academies Press, 2013.

[2]C. Shannon, "A mathematical theory of communication," *Bell System Tech. J.*, vol. 27, pp. 379–423, 623–656, Jul. 1948.

[3]V. Vapnik, *The nature of statistical learning theory*, 2nd ed. Springer, 2000.

[4]G. I. Hinton, "Machine learning," Y. Kodratoff and R. S. Michalski, Eds., 1990, ch. Connectionist Learning Procedures, pp. 555–610.

[5]H. B. Barlow, "Unsupervised learning," *Neural Comput.*, vol. 1, no. 3, pp. 295–311, 1989.

[6]A. Pouget, J. M. Beck, W. J. Ma, and P. E. Latham, "Probabilistic brains: knowns and unknowns," *Nat. Neurosci.*, vol. 16, no. 9, pp. 1170–1178, 2013.

[7]H. Barlow, "The exploitation of regularities in the environment by the brain," *Behav. Brain Sci.*, vol. 24, pp. 602–607, Aug. 2001.

[8]Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[9]Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.

[10]A. R. Barron, "Approximation and estimation bounds for artificial neural networks," *Machine Learning*, vol. 14, no. 1, pp. 115–133, 1994.

[11]J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[12]A. R. Barron and T. M. Cover, "Minimum complexity density estimation," *IEEE Trans. Inf. Theory*, vol. 37, no. 4, pp. 1034–1054, Jul 1991.

[13]S. Boucheron, O. Bousquet, and G. Lugosi, "Theory of classification: a survey of some recent advances," *ESAIM: Probability and Statistics*, vol. 9, pp. 323–375, 2005.

[14]N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *J. of Mach. Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[15]A. Achille and S. Soatto, "Information Dropout: learning optimal representations through noisy computation," *ArXiv e-prints*, Nov. 2016.

[16]D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. of the 2nd Int. Conf. on Learning Representations (ICLR)*, 2013.

[17]C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," *CoRR*, vol. abs/1611.03530, 2016.

[18]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.

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

[20]N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. of the 37th Annu. Allerton Conf. on Communication, Control and Computing*, 1999, pp. 368–377.

[21]D. Russo and J. Zou, "How much does your data exploration overfit? controlling bias via information usage," *arXiv:1511.05219 [cs, stat]*, Nov 2015, arXiv: 1511.05219.

[22]A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 2524–2533.

[23]A. Achille and S. Soatto, "Emergence of invariance and disentangling in deep representations," *arXiv:1706.01350 [cs, stat]*, Jun 2017, arXiv: 1706.01350.

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

[25]V. N. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, September 1998.

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

[27]C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE Nat. Conv. Rec*, vol. 4, no. 142-163, p. 1, 1959.

[28]R. Dobrushin and B. Tsybakov, "Information transmission with additional noise," *IEEE Trans. Inf. Theory*, vol. 8, no. 5, pp. 293–304, September 1962.

[29]T. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," *IEEE Trans. Inf. Theory*, vol. 60, no. 1, pp. 740–761, 2014.

[30]M. Vera, L. R. Vega, and P. Piantanida, "Collaborative representation learning," *ArXiv e-prints*, 2016.

[31]N. Slonim and N. Tishby, "Document clustering using word clusters via the information bottleneck method," in *Proc. of the 23rd Annu. Int. ACM SIGIR Conf. on Research and Develop. in Inform. Retrieval*. ACM, 2000, pp. 208–215.

[32]L. Wang, M. Chen, M. Rodrigues, D. Wilcox, R. Calderbank, and L. Carin, "Information-theoretic compressive measurement design," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1150–1164, Jun 2017.

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

[34]M. Vera, L. R. Vega, and P. Piantanida, "Compression-based regularization with an application to multi-task learning," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 12, pp. 1063–1076, 2018.

[35]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.

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

[37]A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," *CoRR*, vol. abs/1612.00410, 2016.

[38]J. Rissanen, "Paper: Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, Sep. 1978.

[39]P. D. Grnwald, I. J. Myung, and M. A. Pitt, *Advances in Minimum Description Length: Theory and Applications (Neural Information Processing)*. The MIT Press, 2005.

[40]S. Arimoto, "On the converse to the coding theorem for discrete memoryless channels (corresp.)," *IEEE Trans. Inf. Theory*, vol. 19, no. 3, pp. 357–359, May 1973.

[41]Y. M. Shtarkov, "Universal sequential coding of single messages," *Problems of Information Transmission*, vol. 23, no. 3, pp. 175–186, 1987.

[42]A. B. Tsybakov, *Introduction to Nonparametric Estimation*, 1st ed. Springer Publishing Company, Incorporated, 2008.

[43]D. Tebbe and S. Dwyer, "Uncertainty and the probability of error (corresp.)," *IEEE Trans. Inf. Theory*, vol. 14, no. 3, pp. 516–518, May 1968.