
PAC-Bayesian Generalization Bound for Density Estimation with Application to Co-clustering

Yevgeny Seldin

School of Computer Science and Engineering
The Hebrew University of Jerusalem, Israel
{seldin,tishby}@cs.huji.ac.il

Naftali Tishby

Abstract

We derive a PAC-Bayesian generalization bound for density estimation. Similar to the PAC-Bayesian generalization bound for classification, the result has the appealingly simple form of a tradeoff between empirical performance and the KL-divergence of the posterior from the prior. Moreover, the PAC-Bayesian generalization bound for classification can be derived as a special case of the bound for density estimation.

To illustrate a possible application of our bound we derive a generalization bound for co-clustering. The bound provides a criterion to evaluate the ability of co-clustering to predict new co-occurrences, thus introducing the notion of generalization to this traditionally unsupervised task.

1 Introduction

The ability to generalize from a sample is one of the key measures of success in machine learning. However, the notion of generalization is commonly associated with supervised learning. One of the main messages of this work is that it is possible to define and analyze generalization properties of unsupervised learning approaches. Similar to supervised learning, optimization of the generalization abilities of an unsupervised learning algorithm can prevent it from overfitting.

We derive a PAC-Bayesian generalization bound for density estimation, which is a typical example of an unsupervised learning task. PAC-Bayesian generalization bounds (McAllester, 1999) are a state-of-the-art

framework for deriving generalization bounds for classification models. They combine simplicity with explicit dependence on model parameters, thus making them easy to use for classifier optimization.

Theorem 1 (PAC-Bayesian bound for classification). *For a hypothesis set \mathcal{H} , a prior distribution P over \mathcal{H} and a loss function L bounded by 1, with a probability greater than $1 - \delta$ over drawing a sample of size N , for all distributions Q over \mathcal{H} :*

$$D(\hat{L}(Q) \| L(Q)) \leq \frac{D(Q \| P) + \frac{1}{2} \ln(4N) - \ln \delta}{N}, \quad (1)$$

where $L(Q) = \mathbb{E}_{Q(h)} L(h)$ is the expected and $\hat{L}(Q) = \mathbb{E}_{Q(h)} \hat{L}(h)$ is the empirical loss of a randomized classifier that draws $h \in \mathcal{H}$ according to Q and then applies h to classify a sample; $D(Q \| P) = \mathbb{E}_{Q(h)} \ln \frac{Q(h)}{P(h)}$ is the KL-divergence between Q and P and $D(\hat{L}(Q) \| L(Q))$ is the KL-divergence between two Bernoulli variables with biases $\hat{L}(Q)$ and $L(Q)$.

Theorem 1 provides $\frac{1}{2} \ln(N)$ improvement over the original formulation in (McAllester, 1999) and was given a simplified proof in (Maurer, 2004). Successful applications of the theorem include a generalization bound for SVMs (Langford, 2005) and a generalization bound for classification by categorical features (Seldin and Tishby, 2008).

In this work we adapt Maurer's (2004) proof to derive a generalization bound for density estimation, which has a form similar to (1). The bound suggests that a good density estimator should optimize a tradeoff between the likelihood of the empirical data (measured as empirical entropy) and the complexity of the estimator (measured as its KL-divergence from a prior). We then show that the PAC-Bayesian bound (1) follows as a special case of the bound for density estimation, thus supporting the wider generality of the latter.

By contrast to the dominant role of generalization in regularization of classification models, a much more

Appearing in Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5. Copyright 2009 by the authors.

popular regularization approach in density estimation is the MDL principle (Grünwald, 2007). However, it is known that MDL is prone to overfitting (Kearns et al., 1997). Barron and Cover (1991) analyzed generalization properties of density estimators through an index of resolvability. The advantage of the PAC-Bayesian bound over the index of resolvability lies in its explicit dependence on model parameters, which makes it easier to use in practice. Devroye et al. (1996) and Devroye and Lugosi (2001) provide an extensive analysis of density estimation from finite samples in the context of the PAC model. The advantage of PAC-Bayesian bounds over the PAC approach lies in their built-in ability to handle heterogeneous and hierarchical model classes such as separating hyperplanes with all possible margins or the grid clustering models analyzed here.

To illustrate an application of the PAC-Bayesian bound for density estimation we derive a generalization bound for co-clustering. The most well-known application of co-clustering is an analysis of word-document co-occurrence matrices (Slonim and Tishby, 2000; Dhillon et al., 2003). We stress that unlike in the traditional formulation, where co-clustering is aimed at approximating the data at hand, we focus on its out-of-sample performance. Namely, we assume that words and documents (two categorical variables, X_1 and X_2) are drawn from some unknown joint probability distribution $p(X_1, X_2)$ and we are given a sample of size N from that distribution. Our goal is to output an estimator $q(X_1, X_2)$ that will be able to predict new co-occurrences generated by p . In practice we can validate a solution by holding out a random subset of co-occurrence events during training and at the end test the log likelihood of the holdout set.

Our estimator $q(X_1, X_2)$ for $p(X_1, X_2)$ is based on a grid clustering model which draws on work by Seldin and Tishby (2008), who used it for classification. By contrast to co-occurrence data analysis, in the classification scenario the entries of a matrix are unknown functions of the parameters (e.g., in the context of collaborative filtering the entries are ratings given by the viewers to movies). In co-occurrence matrices, however, the entries are joint probability distributions of the parameters. Although it was shown that both problems can be solved within a unified framework (Banerjee et al., 2007) they are different in nature. This can easily be seen by noting that if we add more viewers and movies to a collaborative filtering matrix the previously observed ratings will not change. However, if we add new words and documents to a word-document co-occurrence matrix we have to renormalize the joint probability distribution and thus change the existing values. This difference in nature is also expressed in the proofs of the generalization bounds.

2 The Law of Large Numbers

We first analyze the rate of convergence of empirical distributions over finite domains around their true values. The following result is based on the method of types in information theory (Cover and Thomas, 1991).

Theorem 2. *Let X_1, \dots, X_N be i.i.d. distributed by $p(X)$ and let $|X|$ be the cardinality of X . Denote by $\hat{p}(X)$ the empirical distribution of X_1, \dots, X_N . Then:*

$$\mathbb{E}e^{ND(\hat{p}(X)||p(X))} \leq (N+1)^{|X|-1}. \quad (2)$$

Proof. Enumerate the possible values of X by $1, \dots, |X|$ and let n_i count the number of occurrences of value i . Let p_i denote the probability of value i and $\hat{p}_i = \frac{n_i}{N}$ be its empirical counterpart. Let $D(\hat{p}||p)$ be a shortcut for $D(\hat{p}(X)||p(X))$ and $H(\hat{p}) = -\sum_i \hat{p}_i \ln \hat{p}_i$ be the empirical entropy. Then:

$$\begin{aligned} \mathbb{E}e^{ND(\hat{p}||p)} &= \sum_{\substack{n_1, \dots, n_{|X|}: \\ \sum_i n_i = N}} \binom{N}{n_1, \dots, n_{|X|}} \cdot \prod_{i=1}^{|X|} p_i^{N\hat{p}_i} \cdot e^{ND(\hat{p}||p)} \\ &\leq \sum_{\substack{n_1, \dots, n_{|X|}: \\ \sum_i n_i = N}} e^{NH(\hat{p})} \cdot e^{N\sum_i \hat{p}_i \ln p_i} \cdot e^{ND(\hat{p}||p)} \\ &= \sum_{\substack{n_1, \dots, n_{|X|}: \\ \sum_i n_i = N}} 1 = \binom{N+|X|-1}{|X|-1} \leq (N+1)^{|X|-1}. \end{aligned} \quad (3)$$

In (3) we use the $\binom{N}{n_1, \dots, n_{|X|}} \leq e^{NH(\hat{p})}$ bound on the multinomial coefficient, which counts the number of sequences with a fixed cardinality profile (type) $n_1, \dots, n_{|X|}$ (Cover and Thomas, 1991). In the second equality in (4) the number of ways to choose n_i -s equals the number of ways we can place $|X|-1$ ones in a sequence of $N+|X|-1$ ones and zeros, where ones symbolize a partition of zeros (“balls”) into $|X|$ bins. \square

It is straightforward to recover theorem 12.2.1 in (Cover and Thomas, 1991) from theorem 2. We even suggest a small improvement over it:

Theorem 3 (12.2.1 in Cover and Thomas, 1991). *Under the notations of theorem 2:*

$$P\{D(\hat{p}(X)||p(X)) \geq \varepsilon\} \leq e^{-N\varepsilon + (|X|-1)\ln(N+1)}. \quad (5)$$

Or, equivalently, with a probability greater than $1 - \delta$:

$$D(\hat{p}(X)||p(X)) \leq \frac{(|X|-1)\ln(N+1) - \ln \delta}{N}. \quad (6)$$

Proof. By Markov's inequality and theorem 2:

$$\begin{aligned} P\{D(\hat{p}\|p) \geq \varepsilon\} &= P\{e^{ND(\hat{p}\|p)} \geq e^{N\varepsilon}\} \\ &\leq \frac{\mathbb{E}e^{ND(\hat{p}\|p)}}{e^{N\varepsilon}} \leq \frac{(N+1)^{|X|-1}}{e^{N\varepsilon}} = e^{-N\varepsilon + (|X|-1)\ln(N+1)}. \end{aligned}$$

□

3 A PAC-Bayesian Generalization Bound for Density Estimation

We extend the results of the previous section by considering a family \mathcal{H} of probability distributions over a common domain. The simplest example is to think of \mathcal{H} consisting of two biased dice. For any distribution Q over \mathcal{H} we define $p_Q(X) = \sum_{h \in \mathcal{H}} Q(h)p_h(X)$, where $p_h(X)$ is a distribution over X induced by the hypothesis h (e.g., one of the dice). It is like choosing to roll one die at random according to Q and repeating the experiment several times. In the next theorem we bound the rate of convergence of the empirical distribution $\hat{p}_Q(X)$ of the above process around $p_Q(X)$. The important point is that the bound holds simultaneously for all distributions Q over \mathcal{H} .

The proof of the following theorem reveals a close relation between the PAC-Bayesian bounds and results in information theory, such as Sanov's theorem (Cover and Thomas, 1991). In addition, we recover the original PAC-Bayesian bound for classification (1) as a special case of the following theorem.

Theorem 4. *Let \mathcal{H} be a hypothesis class (possibly uncountably infinite), such that each member h of \mathcal{H} induces a probability distribution $p_h(X)$ over a random variable X with cardinality $|X|$. Let P be a prior distribution over \mathcal{H} . Let Q be an arbitrary distribution over \mathcal{H} and $p_Q(X) = \mathbb{E}_{Q(h)}p_h(X)$ a distribution over X induced by a randomized process that first chooses a hypothesis $h \in \mathcal{H}$ according to $Q(h)$ and then draws X according to $p_h(X)$. Let $\hat{p}_Q(X)$ be an empirical distribution over X obtained by performing N draws of X according to $p_Q(X)$. Let $D(\hat{p}_Q\|p_Q)$ be a shortcut for $D(\hat{p}_Q(X)\|p_Q(X))$. Then for all distributions Q over \mathcal{H} with a probability greater than $1 - \delta$:*

$$D(\hat{p}_Q\|p_Q) \leq \frac{D(Q\|P) + (|X| - 1)\ln(N + 1) - \ln \delta}{N}. \quad (7)$$

Proof. We adapt Maurer's (2004) proof of theorem 1 by substituting the result on the concentration of $\mathbb{E}e^{ND(\hat{L}(Q)\|L(Q))}$ in the classification scenario with the result on the concentration of $\mathbb{E}e^{ND(\hat{p}(X)\|p(X))}$ in theorem 2. Denoting the sample by $S = \{X_1, \dots, X_N\}$, by (2) we have:

$$(N + 1)^{|X|-1} \geq \mathbb{E}_{P(h)}\mathbb{E}_S e^{ND(\hat{p}_h(X)\|p_h(X))}$$

$$\begin{aligned} &= \mathbb{E}_S \mathbb{E}_{P(h)} e^{ND(\hat{p}_h(X)\|p_h(X))} \\ &= \mathbb{E}_S \mathbb{E}_{Q(h)} e^{ND(\hat{p}_h(X)\|p_h(X)) - \ln \frac{P(h)}{Q(h)}} \\ &\geq \mathbb{E}_S e^{\mathbb{E}_{Q(h)}[ND(\hat{p}_h(X)\|p_h(X)) - \ln \frac{P(h)}{Q(h)}]} \\ &\geq \mathbb{E}_S e^{ND(\hat{p}_Q(X)\|p_Q(X)) - D(Q\|P)}, \end{aligned} \quad (8)$$

$$\geq \mathbb{E}_S e^{ND(\hat{p}_Q(X)\|p_Q(X)) - D(Q\|P)}, \quad (9)$$

where (8) is justified by the convexity of the exponent function and (9) by the convexity of the Kullback-Leibler divergence. By (9) and Markov's inequality:

$$\begin{aligned} P_S\{D(\hat{p}_Q(X)\|p_Q(X)) > \varepsilon\} &= \\ &= P_S\{e^{ND(\hat{p}_Q(X)\|p_Q(X)) - D(Q\|P)} > e^{N\varepsilon - D(Q\|P)}\} \\ &\leq \frac{(N + 1)^{|X|-1}}{e^{N\varepsilon - D(Q\|P)}}. \end{aligned} \quad (10)$$

By choosing ε so that the right-hand side of (10) is bounded by δ we obtain (7). □

To recover the PAC-Bayesian theorem 1 from theorem 4 in the case of zero-one loss let X be the error variable. Then $L(h) = \mathbb{E}X = p_h\{X = 1\}$ and $L(Q) = p_Q\{X = 1\}$. As well, $|X| = 2$. Substituting this into (7) we obtain (1) up to a factor of $\frac{1}{2} \ln N$. By the convexity of $D(\hat{L}(Q)\|L(Q))$ it is possible to show that the result holds for any loss bounded by 1 (Maurer, 2004). The improvement of $\frac{1}{2} \ln N$ in theorem 1 is achieved by showing that in the case where $L(h) = \mathbb{E}X$ the expectation $\mathbb{E}e^{ND(\hat{L}(h)\|L(h))} \leq 2\sqrt{N}$ instead of the more general bound $\mathbb{E}e^{ND(\hat{p}(X)\|p(X))} \leq (N + 1)^{|X|-1}$ for distributions we have in theorem 2.

4 Smoothing

Although we bounded $D(\hat{p}_Q(X)\|p_Q(X))$ in the previous section, the value of interest in many applications is usually not $D(\hat{p}_Q(X)\|p_Q(X))$, but rather $-\mathbb{E}_{p_Q(X)} \ln \hat{p}_Q(X)$. The latter corresponds to the expected performance (in log loss) of the model $\hat{p}_Q(X)$ when samples are drawn by $p_Q(X)$. Or, in the language of information theory, to the expected code length of encoder \hat{p}_Q when samples are generated by p_Q . This follows the classical PAC spirit, which states that performance guarantees should be provided with respect to the true, unknown data-generating distribution. Unfortunately, $-\mathbb{E}_{p_Q(X)} \ln \hat{p}_Q(X)$ cannot be bounded since $\hat{p}_Q(X)$ is not bounded from zero. To cope with this, we define a smoothed version of \hat{p} we call q :

$$q_h(X) = \frac{\hat{p}_h(X) + \gamma}{1 + \gamma|X|}, \quad (11)$$

$$q_Q(X) = \mathbb{E}_{Q(h)} q_h(X) = \frac{\hat{p}_Q(X) + \gamma}{1 + \gamma|X|}. \quad (12)$$

In the following theorem we show that if $D(\hat{p}_Q(X)||p_Q(X)) \leq \varepsilon(Q)$ and $\gamma = \frac{\sqrt{\varepsilon(Q)/2}}{|X|}$, then $-\mathbb{E}_{p_Q(X)} \ln q_Q(X)$ is roughly within $\pm \sqrt{\varepsilon(Q)/2} \ln |X|$ range around $H(\hat{p}_Q(X))$. The bound on $D(\hat{p}_Q(X)||p_Q(X))$ is naturally obtained by theorem 4. Thus, the performance of the density estimator q_Q is optimized by distribution Q that minimizes the tradeoff between $H(\hat{p}_Q(X))$ and $\frac{1}{N}D(Q||P)$.

Note that for a uniform distribution $u(X) = \frac{1}{|X|}$ the value of $-\mathbb{E}_{p(X)} \ln u(X) = \ln |X|$. Thus, the theorem is interesting when $\sqrt{\varepsilon(Q)/2}$ is significantly smaller than 1. For technical reasons we encounter in the proofs of the next section, the upper bound in the following theorem is stated for $-\mathbb{E}_{p_Q(X)} \ln q_Q(X)$ and for $-\mathbb{E}_{Q(h)} \mathbb{E}_{p_h(X)} \ln q_h(X)$. We also denote $\varepsilon = \varepsilon(Q)$ for brevity.

Theorem 5. *Let X be a random variable distributed according to $p_Q(X)$ and assume that $D(\hat{p}_Q(X)||p_Q(X)) \leq \varepsilon$. Then $-\mathbb{E}_{p_Q(X)} \ln q_Q(X)$ is minimized for $\gamma = \frac{\sqrt{\varepsilon/2}}{|X|}$. For this value of γ the following inequalities hold:*

$$\begin{aligned} -\mathbb{E}_{p_Q(X)} \ln q_Q(X) &\leq -\mathbb{E}_{Q(h)} \mathbb{E}_{p_h(X)} \ln q_h(X) & (13) \\ &\leq H(\hat{p}_Q(X)) + \sqrt{\varepsilon/2} \ln |X| + \phi(\varepsilon), & (14) \end{aligned}$$

$$-\mathbb{E}_{p_Q(X)} \ln q_Q(X) \geq H(\hat{p}_Q(X)) - \sqrt{\varepsilon/2} \ln |X| - \psi(\varepsilon), \quad (15)$$

where:

$$\psi(\varepsilon) = \sqrt{\frac{\varepsilon}{2}} \ln \frac{1 + \sqrt{\frac{\varepsilon}{2}}}{\sqrt{\frac{\varepsilon}{2}}} \quad \text{and} \quad \phi(\varepsilon) = \psi(\varepsilon) + \ln(1 + \sqrt{\frac{\varepsilon}{2}}).$$

Note that both $\phi(\varepsilon)$ and $\psi(\varepsilon)$ go to zero approximately as $-\sqrt{\varepsilon/2} \ln \sqrt{\varepsilon/2}$.

Proof. By the KL-divergence bound on the L_1 norm (Cover and Thomas, 1991):

$$\|\hat{p}_Q(X) - p_Q(X)\|_1 \leq \sqrt{2D(\hat{p}_Q(X)||p_Q(X))} \leq \sqrt{2\varepsilon}. \quad (16)$$

We start with a proof of (14):

$$\begin{aligned} &-\mathbb{E}_{Q(h)} \mathbb{E}_{p_h(X)} \ln q_h(X) \\ &= \mathbb{E}_{Q(h)} \mathbb{E}_{[\hat{p}_h(X) - p_h(X)]} \ln q_h(X) - \mathbb{E}_{Q(h)} \mathbb{E}_{\hat{p}_h(X)} \ln q_h(X) \\ &= \mathbb{E}_{Q(h)} \mathbb{E}_{[\hat{p}_h(X) - p_h(X)]} \ln \frac{\hat{p}_h(X) + \gamma}{1 + \gamma|X|} \\ &\quad - \mathbb{E}_{Q(h)} \mathbb{E}_{\hat{p}_h(X)} \ln \frac{\hat{p}_h(X) + \gamma}{1 + \gamma|X|} \\ &\leq -\frac{1}{2} \|\hat{p}_Q(X) - p_Q(X)\|_1 \ln \frac{\gamma}{1 + \gamma|X|} \end{aligned}$$

$$\begin{aligned} &+ \mathbb{E}_{Q(h)} H(\hat{p}_h(X)) + \ln(1 + \gamma|X|) \\ &\leq H(\hat{p}_Q(X)) - \sqrt{\varepsilon/2} \ln \frac{\gamma}{1 + \gamma|X|} + \ln(1 + \gamma|X|), \end{aligned} \quad (17)$$

where (17) is justified by the concavity of the entropy function H and (16). By differentiation (17) is minimized by $\gamma = \frac{\sqrt{\varepsilon/2}}{|X|}$. By substitution of this value of γ into (17) we obtain (14). Inequality (13) is justified by the concavity of the \ln function. Finally, we prove the lower bound (15):

$$\begin{aligned} &-\mathbb{E}_{p_Q(X)} \ln q_Q(X) \\ &= \sum_x (\hat{p}_Q(x) - p_Q(x)) \ln q_Q(x) - \sum_x \hat{p}_Q(x) \ln q_Q(x) \\ &\geq -\frac{1}{2} \|\hat{p}_Q(X) - p_Q(X)\|_1 \ln \frac{1 + \gamma|X|}{\gamma} + H(\hat{p}_Q(X)) \\ &\geq H(\hat{p}_Q(X)) - \sqrt{\varepsilon/2} \ln \frac{|X|(1 + \sqrt{\varepsilon/2})}{\sqrt{\varepsilon/2}}. \end{aligned}$$

□

5 Generalization Bound for Density Estimation with Grid Clustering

As an example of an application of the bounds developed in the previous sections we derive a generalization bound for density estimation with grid clustering. The goal is to find a good estimator for an unknown joint probability distribution $p(X_1, \dots, X_d)$ over a d -dimensional product space $\mathcal{X}_1 \times \dots \times \mathcal{X}_d$ based on a sample of size N from p . Each of the i sub-spaces, \mathcal{X}_i , is assumed to be a categorical space of known cardinality n_i . As an illustrative example, think of estimating a joint probability of words and documents (X_1 and X_2) from their co-occurrence matrix. We denote elements of \mathcal{X}_i by x_i and random variables accepting values in \mathcal{X}_i by X_i . The goodness of an estimator q for p is measured as $-\mathbb{E}_{p(X_1, \dots, X_d)} \ln q(X_1, \dots, X_d)$.

By theorem 3, to obtain a meaningful bound for a direct estimation of $p(X_1, \dots, X_d)$ we need N to be exponential in n_i -s, since the cardinality of the random variable $\langle X_1, \dots, X_d \rangle$ is $\prod_i n_i$. To reduce this dependency to be linear in $\sum_i n_i$ we restrict the estimator $q(X_1, \dots, X_d)$ to be of the factor form:

$$\begin{aligned} q(X_1, \dots, X_d) &= \sum_{C_1, \dots, C_d} q(C_1, \dots, C_d) \prod_i q(X_i | C_i) \\ &= \sum_{C_1, \dots, C_d} q(C_1, \dots, C_d) \prod_i \frac{q(X_i)}{q(C_i)} q(C_i | X_i). \end{aligned} \quad (18)$$

We emphasize that the above decomposition assumption is only on the estimator q , but not on the generating distribution p .

5.1 Hypothesis Space

In order to apply the PAC-Bayesian bound we have to define a hypothesis space \mathcal{H} . The hypothesis space \mathcal{H} we choose is a set of all hard grid partitions of the input product space $\mathcal{X}_1 \times \dots \times \mathcal{X}_d$. A partition is defined by mapping each $x_i \in \mathcal{X}_i$ to a cluster variable C_i . Thus, in \mathcal{H} every x_i is mapped to a single C_i and soft assignment is achieved by averaging over hard assignments. The set of distributions $Q = \{q(C_i|X_i)\}_{i=1}^d$, the rightmost in (18), define a distribution over \mathcal{H} .

We note that for any distribution $p(X_1, \dots, X_d)$ the distribution Q over \mathcal{H} induces a distribution $p_Q(C_1, \dots, C_d)$, which is a probability distribution over the coarsened space defined by cluster variables C_1, \dots, C_d and corresponding to p - see (21) below. In section 5.3 we provide rates of convergence of empirical frequencies $\hat{p}_Q(C_1, \dots, C_d)$, defined by the empirical distribution $\hat{p}(X_1, \dots, X_d)$ and Q , toward the true ones. The rates depend on the size of the coarsened cluster space rather than on the size of the input space. In section 5.4 we show how to smooth the empirical counts in order to obtain an estimator $q(X_1, \dots, X_d)$ of the form (18) for p with guarantees on $-\mathbb{E}_{p(X_1, \dots, X_d)} \ln q(X_1, \dots, X_d)$.

5.2 Some Additional Definitions

We denote the cardinality of C_i by m_i . The value of m_i can vary in the range of $1 \leq m_i \leq n_i$, where $m_i = 1$ corresponds to the case where all x_i -s are mapped to a single cluster and $m_i = n_i$ corresponds to the case where each x_i is mapped to a separate cluster. For a fixed distribution Q over \mathcal{H} the values of m_i -s are fixed. We use this to denote the number of partition cells in hypotheses selected by Q by $M = \prod_i m_i$.

We define $q_i(C_i) = \frac{1}{n_i} \sum_{x_i} q(C_i|x_i)$ to be a marginal distribution over C_i corresponding to a *uniform* distribution over \mathcal{X}_i and the conditional distribution $q(C_i|X_i)$ of our choice. Then $I_U(X_i; C_i) = \frac{1}{n_i} \sum_{x_i, c_i} q(c_i|x_i) \ln[q(c_i|x_i)/q_i(c_i)]$ is the mutual information between X_i and C_i corresponding to the uniform distribution over \mathcal{X}_i and the conditional distribution $q(C_i|X_i)$ of our choice.

Let $h \in \mathcal{H}$ be a hard clustering and let $c_i = h(x_i)$ denote the cluster that x_i is mapped to in h . We define the distribution over $\langle C_1, \dots, C_d \rangle$ induced by p and h , its extension for Q , and the corresponding marginals:

$$p_h(c_1, \dots, c_d) = \sum_{\substack{x_1, \dots, x_d: \\ h(x_i)=c_i}} p(x_1, \dots, x_d), \quad (19)$$

$$p_h(c_i) = \sum_{x_i: h(x_i)=c_i} p(x_i), \quad (20)$$

$$\begin{aligned} p_Q(c_1, \dots, c_d) &= \sum_h Q(h) p_h(c_1, \dots, c_d) \\ &= \sum_{x_1, \dots, x_d} p(x_1, \dots, x_d) \prod_i q(c_i|x_i), \end{aligned} \quad (21)$$

$$p_Q(c_i) = \sum_h Q(h) p_h(c_i) = \sum_{x_i} p(x_i) q(c_i|x_i). \quad (22)$$

Recall that we have no access to p , but only to the empirical distribution $\hat{p}(X_1, \dots, X_d)$ defined by the sample. We define the empirical counterparts of p_h and p_Q , that we call \hat{p}_h and \hat{p}_Q , by substitution of $\hat{p}(X_1, \dots, X_d)$ and $\hat{p}(X_i)$ instead of $p(X_1, \dots, X_d)$ and $p(X_i)$ in equations (19), (20), (21), and (22) above.

We extend the definition of p_h to $\langle X_1, \dots, X_d \rangle$ by requiring it to have the factor form of (18):

$$\begin{aligned} p_h(X_1, \dots, X_d) &= p_h(h(X_1), \dots, h(X_d)) \prod_i \frac{p(X_i)}{p_h(h(X_i))} \\ &= p_h(C_1, \dots, C_d) \prod_i \frac{p(X_i)}{p_h(C_i)} \quad \text{for } C_i = h(X_i), \end{aligned} \quad (23)$$

$$p_Q(X_1, \dots, X_d) = \sum_{C_1, \dots, C_d} p_Q(C_1, \dots, C_d) \prod_i \frac{p(X_i)}{p_Q(C_i)} q(C_i|X_i). \quad (24)$$

Finally, we define the empirical counterparts by substitution of \hat{p} instead of p in the equations (23) and (24).

5.3 Assurances on Empirical Density Estimates in Grid Clustering

Our first result in this section concerns convergence (in KL-divergence) of the empirical estimates $\hat{p}_Q(X_1, \dots, X_d)$, $\hat{p}_Q(C_1, \dots, C_d)$, and $\hat{p}(X_i)$ to their true values. An interesting point about the following theorem is that the cardinality of the random variable $\langle X_1, \dots, X_n \rangle$ is $\prod_i n_i$. A direct application of theorem 4 would insert a dependency on $\prod_i n_i$ into the bound. However, by using the factor form of $p_h(X_1, \dots, X_d)$ we are able to reduce the dependency to $M + \sum_i n_i$, which is linear instead of exponential in n_i - see (27) below. The bounds on $\hat{p}_Q(C_1, \dots, C_d)$ and $\hat{p}(X_i)$ in (25) and (26) are used in the next section to construct an estimator for p with generalization guarantees.

Theorem 6. *For any probability measure p over instances and an i.i.d. sample S of size N according to p , with a probability of at least $1 - \delta$ for all grid clusterings $Q = \{q_i(C_i|X_i)\}_{i=1}^d$ the following holds simultaneously:*

$$D(\hat{p}_Q(C_1, \dots, C_d) \| p_Q(C_1, \dots, C_d)) \leq \frac{\sum_i n_i I_U(X_i; C_i) + K_1}{N} \quad (25)$$

$$K_1 = \sum_i m_i \ln n_i + (M - 1) \ln(N + 1) + \ln \frac{d+1}{\delta},$$

$$D(\hat{p}(X_i) \| p(X_i)) \leq \frac{(n_i - 1) \ln(N + 1) + \ln \frac{d+1}{\delta}}{N}. \quad (26)$$

As well, with a probability greater than $1 - \delta$:

$$D(\hat{p}_Q(X_1, \dots, X_d) \| p_Q(X_1, \dots, X_d)) \leq \frac{\sum_i n_i I_U(X_i; C_i) + K_2}{N} \quad (27)$$

$$K_2 = \sum_i m_i \ln n_i + \left[M + \sum_i n_i - d - 1 \right] \ln(N + 1) - \ln \delta.$$

Proof. The proof is based on theorem 4. To apply the theorem we have to define a prior P over \mathcal{H} and then to calculate $D(Q \| P)$. There is a small caveat, since the cardinality $M = \prod_i m_i$ of the random variable $\langle C_1, \dots, C_d \rangle$ changes when we change m_i -s. However, we recall that for a fixed Q the values of m_i -s are fixed. Let us denote by $\bar{m} = (m_1, \dots, m_d)$ a vector counting the number of clusters used along each dimension. We slice \mathcal{H} into slices denoted by $\mathcal{H}_{\bar{m}}$, according to the number of clusters used along each dimension. Obviously, $\mathcal{H}_{\bar{m}}$ -s are disjoint. We handle each $\mathcal{H}_{\bar{m}}$ independently and then combine the results to obtain theorem 6. We use $h|_i$ to denote a partition induced by h along dimension i , thus $h = h|_1 \times \dots \times h|_d$. In the subsequent two lemmas, adapted from (Seldin and Tishby, 2008) with some improvements, we define a prior P over $\mathcal{H}_{\bar{m}}$ and then calculate $D(Q \| P)$.

Lemma 1. *It is possible to define a prior P over $\mathcal{H}_{\bar{m}}$ that satisfies:*

$$P(h) \geq \frac{1}{\exp \left[\sum_i (n_i H(q_{h|_i}) + (m_i - 1) \ln n_i) \right]}, \quad (28)$$

where $q_{h|_i}$ denotes the cardinality profile of cluster sizes along dimension i of a partition corresponding to h .

Lemma 2. *For the prior defined in lemma 1:*

$$D(Q \| P) \leq \sum_i [n_i I_U(X_i; C_i) + (m_i - 1) \ln n_i]. \quad (29)$$

Proof of Lemma 1. To define the prior P over $\mathcal{H}_{\bar{m}}$ we count the hypotheses in $\mathcal{H}_{\bar{m}}$. There are $\binom{n_i-1}{m_i-1} \leq n_i^{m_i-1}$ possibilities to choose a cluster cardinality profile along a dimension i . (Each of the m_i clusters has a size of at least one. To define a cardinality profile we are free to distribute the “excess mass” of $n_i - m_i$ among the m_i clusters. The number of possible distributions equals the number of possibilities to place $m_i - 1$ ones in a sequence of $(n_i - m_i) + (m_i - 1) = n_i - 1$ ones and zeros.) For a fixed cardinality profile $|c_{i1}|, \dots, |c_{im_i}|$ (over a single dimension) there are

$\binom{n_i}{|c_{i1}|, \dots, |c_{im_i}|} \leq e^{n_i H_U(C_i)}$ possibilities to assign X_i -s to the clusters. Putting all the combinatorial calculations together we can define a distribution $P(h)$ over $\mathcal{H}_{\bar{m}}$ that satisfies (28). \square

At this juncture it is worth stressing that by contrast to most applications of the PAC-Bayesian bound, in our case the prior P and the posterior Q are defined over slightly different hypothesis spaces. The posterior Q is defined for *named* clusterings - we explicitly specify for each X_i the “name” of C_i it is mapped to. And the prior P is defined over *unnamed* partitions - we only check the cardinality profile of C_i , but we cannot recover which X_i -s are mapped to a given C_i . Nevertheless, the “named” distribution Q induces a distribution over the “unnamed” space by summing over all possible name permutations. This enables us to compute $D(Q \| P)$ we need for the bound.

Proof of Lemma 2. We use the decomposition $D(Q \| P) = -\mathbb{E}_Q P(h) - H(Q)$ and bound $-\mathbb{E}_Q P(h)$ and $H(Q)$ separately. We further decompose $P(h) = P(h|_1) \dots P(h|_d)$ and $Q(h)$ in a similar manner. Then $-\mathbb{E}_Q \ln P(h) = -\sum_i \mathbb{E}_Q \ln P(h|_i)$, and similarly for $D(Q \| P)$. Therefore, we can treat each dimension independently.

To bound $-\mathbb{E}_Q \ln P(h|_i)$ recall that $q_i(c_i) = \frac{1}{n_i} \sum_{x_i} q(c_i | x_i)$ is the expected distribution over cardinalities of clusters along dimension i if we draw a cluster c_i for each value x_i of \mathcal{X}_i according to $q(C_i | X_i)$. Let $q_{h|_i}$ be a cluster cardinality profile obtained by such an assignment and corresponding to a hypothesis $h|_i$. Then by lemma 1:

$$-\mathbb{E}_Q \ln P(h|_i) \leq (m_i - 1) \ln n_i + n_i \mathbb{E}_{q_i} H(q_{h|_i}). \quad (30)$$

To bound $\mathbb{E}_{q_i} H(q_{h|_i})$ we use the result on the negative bias of empirical entropy estimates cited below. See (Paninski, 2003) for a proof.

Theorem 7 (Paninski, 2003). *Let X_1, \dots, X_N be i.i.d. distributed by $p(X)$ and let $\hat{p}(X)$ be their empirical distribution. Then:*

$$\mathbb{E}_p H(\hat{p}) = H(p) - \mathbb{E}_p D(\hat{p} \| p) \leq H(p). \quad (31)$$

By (31) $\mathbb{E}_{q_i} H(q_{h|_i}) \leq H(q_i)$. Substituting this into (30) yields:

$$-\mathbb{E}_Q \ln P(h|_i) \leq n_i H(q_i) + (m_i - 1) \ln n_i. \quad (32)$$

Now we turn to compute $\mathbb{E}_Q \ln Q(h|_i)$. To do so we bound $\ln Q(q_{h|_i})$ from above. The bound follows from the fact that if we draw n_i values of C_i according to $q_i(C_i | X_i)$ the probability of the resulting

type is bounded from above by $e^{-n_i H_U(C_i|X_i)}$, where $H_U(C_i|X_i) = -\frac{1}{n_i} \sum_{x_i, c_i} q_i(c_i|x_i) \ln q_i(c_i|x_i)$ (see theorem 12.1.2 in (Cover and Thomas, 1991)). Thus, $\mathbb{E}_Q \ln Q(h|_i) \leq -n_i H_U(C_i|X_i)$, which together with (32) and the identity $I_U(X_i; C_i) = H(q_i) - H_U(C_i|X_i)$ completes the proof of (29). \square

Recall that there are $\prod_i n_i$ disjoint subspaces $\mathcal{H}_{\bar{m}}$ of \mathcal{H} and that each Q is defined over a single $\mathcal{H}_{\bar{m}}$. By theorem 4 and lemma 2, for the prior P over $\mathcal{H}_{\bar{m}}$ defined in lemma 1, with a probability greater than $1 - \frac{\delta}{(d+1)\prod_i n_i}$ we obtain (25) for each $\mathcal{H}_{\bar{m}}$. In addition, by theorem 3 with a probability greater than $1 - \frac{\delta}{d+1}$ inequality (26) holds for each X_i . By a union bound over the $\prod_i n_i$ subspaces of \mathcal{H} and the d variables X_i we obtain that (25) and (26) hold simultaneously for all Q and X_i with a probability greater than $1 - \delta$.

To prove (27), fix some hard partition h and let $C_i^h = h(X_i)$. Then:

$$\begin{aligned} & D(\hat{p}_h(X_1, \dots, X_d) \| p_h(X_1, \dots, X_d)) \\ &= D(\hat{p}_h(X_1, \dots, X_d, C_1^h, \dots, C_d^h) \| p_h(X_1, \dots, X_d, C_1^h, \dots, C_d^h)) \\ &= D(\hat{p}_h(C_1, \dots, C_d) \| p_h(C_1, \dots, C_d)) \\ &\quad + D(\hat{p}_h(X_1, \dots, X_d | C_1^h, \dots, C_d^h) \| p_h(X_1, \dots, X_d | C_1^h, \dots, C_d^h)) \\ &= D(\hat{p}_h(C_1, \dots, C_d) \| p_h(C_1, \dots, C_d)) \\ &\quad + \sum_i D(\hat{p}_h(X_i | C_i^h) \| p_h(X_i | C_i^h)) \\ &= D(\hat{p}_h(C_1, \dots, C_d) \| p_h(C_1, \dots, C_d)) \\ &\quad + \sum_i D(\hat{p}(X_i) \| p(X_i)) - \sum_i D(\hat{p}_h(C_i^h) \| p_h(C_i^h)) \\ &\leq D(\hat{p}_h(C_1, \dots, C_d) \| p_h(C_1, \dots, C_d)) + \sum_i D(\hat{p}(X_i) \| p(X_i)) \end{aligned}$$

And:

$$\begin{aligned} & \mathbb{E} e^{ND(\hat{p}_h(X_1, \dots, X_d) \| p_h(X_1, \dots, X_d))} \\ &\leq \mathbb{E} e^{N[D(\hat{p}_h(C_1, \dots, C_d) \| p_h(C_1, \dots, C_d)) + \sum_i D(\hat{p}_h(X_i) \| p_h(X_i))]} \\ &= \mathbb{E} e^{ND(\hat{p}_h(C_1, \dots, C_d) \| p_h(C_1, \dots, C_d))} \prod_i \mathbb{E} e^{ND(\hat{p}_h(X_i) \| p_h(X_i))} \\ &\leq (N+1)^{M + \sum_i n_i - (d+1)}, \end{aligned}$$

where the last inequality is by theorem 2. From here, following the lines of the proof of theorem 4 we obtain:

$$\begin{aligned} & (N+1)^{M + \sum_i n_i - (d+1)} \\ &\geq \mathbb{E} e^{ND(\hat{p}_Q(X_1, \dots, X_d) \| p_Q(X_1, \dots, X_d)) - D(Q \| P)}, \end{aligned}$$

and, continuing with that proof that with a probability greater than $1 - \delta$:

$$\begin{aligned} & D(\hat{p}_Q(X_1, \dots, X_d) \| p_Q(X_1, \dots, X_d)) \\ &\leq \frac{D(Q \| P) + [M + \sum_i n_i - d - 1] \ln(N+1) - \ln \delta}{N}. \end{aligned}$$

Finally, taking the prior P over \mathcal{H} defined in lemma 1 (this time we give a weight of $(\prod_i n_i)^{-1}$ to each $\mathcal{H}_{\bar{m}}$ and obtain a prior over the whole \mathcal{H}) and taking the calculation of $D(Q \| P)$ in lemma 2 we obtain (27). \square

5.4 Construction of a Density Estimator

Our goal now is to construct an estimation $q(X_1, \dots, X_d)$ for $p(X_1, \dots, X_d)$ with guarantees on performance measured as $-\mathbb{E}_{p(X_1, \dots, X_d)} \ln q(X_1, \dots, X_d)$. Note that although we bounded $D(\hat{p}_Q(X_1, \dots, X_d) \| p_Q(X_1, \dots, X_d))$ it does not yet provide guarantees on the performance of $\hat{p}_Q(X_1, \dots, X_d)$ since it is not bounded from zero. It is also problematic to use theorem 5 to smooth $\hat{p}_Q(X_1, \dots, X_d)$ directly, since the cardinality of the random variable $\langle X_1, \dots, X_d \rangle$ is $\prod_i n_i$ and this factor will enter into smoothing and subsequent bounds. To get around this we utilize the factor form of p_h and the bounds (25) and (26). We define an estimator q in the following way:

$$q_h(C_1, \dots, C_d) = \frac{\hat{p}_h(C_1, \dots, C_d) + \gamma}{1 + \gamma M}, \quad (33)$$

$$q(X_i) = \frac{\hat{p}(X_i) + \gamma_i}{1 + \gamma_i n_i}, \quad (34)$$

$$q_h(c_i) = \sum_{x_i: h(x_i)=c_i} q(x_i), \quad (35)$$

$$q_h(X_1, \dots, X_d) = q_h(C_1^h, \dots, C_d^h) \prod_i \frac{q_h(X_i)}{q_h(C_i^h)}. \quad (36)$$

And for a distribution Q over \mathcal{H} :

$$q_Q(C_1, \dots, C_d) = \frac{\hat{p}_Q(C_1, \dots, C_d) + \gamma}{1 + \gamma M}, \quad (37)$$

$$q_Q(C_i) = \sum_{x_i} q(x_i) q(C_i | x_i) = \frac{\hat{p}_Q(C_i) + \gamma_i q_i(C_i) n_i}{1 + \gamma_i n_i}, \quad (38)$$

$$\begin{aligned} q_Q(X_1, \dots, X_d) &= \sum_h Q(h) q_h(X_1, \dots, X_d) \\ &= \sum_{C_1, \dots, C_d} q_Q(C_1, \dots, C_d) \prod_i \frac{q(X_i)}{q_Q(C_i)} q(C_i | X_i). \end{aligned} \quad (39)$$

In the following theorem we provide a bound on $-\mathbb{E}_{p(X_1, \dots, X_d)} \ln q_Q(X_1, \dots, X_d)$. Note, that we take the expectation with respect to the true, unknown distribution p that may have an arbitrary form.

Theorem 8. *For the density estimator $q_Q(X_1, \dots, X_d)$ defined by equations (34), (37), (38), and (39), $-\mathbb{E}_{p(X_1, \dots, X_d)} \ln q_Q(X_1, \dots, X_d)$ attains its minimum at $\gamma = \frac{\sqrt{\varepsilon_i/2}}{M}$ and $\gamma_i = \frac{\sqrt{\varepsilon_i/2}}{n_i}$, where ε is defined by the right-hand side of (25) and ε_i is defined by the right-hand*

side of (26). At this optimal level of smoothing, with a probability greater than $1 - \delta$ for all grid clusterings Q :

$$\begin{aligned} & -\mathbb{E}_{p(X_1, \dots, X_d)} \ln q_Q(X_1, \dots, X_d) \\ & \leq -I(\hat{p}_Q(C_1, \dots, C_d)) + \sqrt{\varepsilon/2} \ln M + \phi(\varepsilon) + K_3 \end{aligned} \quad (40)$$

$$K_3 = \sum_i [H(\hat{p}(X_i)) + 2\sqrt{\varepsilon_i/2} \ln n_i + \phi(\varepsilon_i) + \psi(\varepsilon_i)],$$

where $I(\hat{p}_Q(C_1, \dots, C_d)) = \sum_i H(\hat{p}_Q(C_i)) - H(\hat{p}_Q(C_1, \dots, C_d))$ is the multi-information between C_1, \dots, C_d with respect to $\hat{p}_Q(C_1, \dots, C_d)$.

Proof.

$$\begin{aligned} & -\mathbb{E}_{p(X_1, \dots, X_d)} \ln q_Q(X_1, \dots, X_d) \\ & = -\mathbb{E}_{p(X_1, \dots, X_d)} \ln \mathbb{E}_{Q(h)} q_h(X_1, \dots, X_d) \\ & \leq -\mathbb{E}_{Q(h)} \mathbb{E}_{p(X_1, \dots, X_d)} \ln q_h(X_1, \dots, X_d) \\ & = -\mathbb{E}_{Q(h)} \mathbb{E}_{p(X_1, \dots, X_d)} \ln q_h(C_1^h, \dots, C_d^h) \prod_i \frac{q(X_i)}{q_h(C_i^h)} \\ & = -\mathbb{E}_{Q(h)} [\mathbb{E}_{p_h(C_1, \dots, C_d)} \ln q_h(C_1, \dots, C_d)] \\ & \quad - \sum_i \mathbb{E}_{p(X_i)} \ln q(X_i) + \sum_i \mathbb{E}_{Q(h)} \mathbb{E}_{p_h(C_i)} \ln q_h(C_i) \\ & \leq -\mathbb{E}_{Q(h)} [\mathbb{E}_{p_h(C_1, \dots, C_d)} \ln q_h(C_1, \dots, C_d)] \\ & \quad - \sum_i \mathbb{E}_{p(X_i)} \ln q(X_i) + \sum_i E_{p_Q(C_i)} \ln q_Q(C_i) \end{aligned}$$

At this point we use (14) to bound the first and the second term and the lower bound (15) to bound the last term and obtain (40). \square

To summarize, recall that $q_Q(X_1, \dots, X_d)$ is defined by our choice of $Q = \{q(C_i|X_i)\}_{i=1}^d$ and quantities determined by Q and the sample $\hat{p}(X_1, \dots, X_d)$. The bound (40) suggests that a good estimator q_Q for $p(X_1, \dots, X_d)$ should optimize a tradeoff between $-I(\hat{p}_Q(C_1, \dots, C_d))$ and $\sum_i \frac{n_i}{N} I_U(X_i; C_i)$. It is easy to see that co-clustering is a special case of our analysis when we have only two variables X_1 and X_2 . The above tradeoff suggests a modification of the original formulation of co-clustering in (Dhillon et al., 2003), which states that a solution should maximize $I(C_1, C_2)$ alone. The suggested tradeoff can be used for model selection.

6 Discussion

We presented a PAC-Bayesian generalization bound for density estimation. The bound suggests a tradeoff between an estimator's complexity and its empirical performance that should be optimized to achieve better out-of-sample performance. We applied the bound

to derive a generalization bound for density estimation with grid clustering in categorical product spaces.

We note that we were able to make use of the factor form of the distribution induced by grid clustering to derive a better concentration result. In future research it would be useful to extend this result to more general independence assumptions and graphical models.

Another direction we would like to highlight is the way we transformed the formulation of co-clustering (an unsupervised task) to make sense of its generalization properties. It would be worthwhile to find a similar transformation of other clustering tasks to conduct their rigorous evaluation.

Acknowledgments: This work was partially supported by Leibnitz Center for Research in Computer Science and the NATO SfP-982480 grant.

References

- Arindam Banerjee, Inderjit Dhillon, Joydeep Ghosh, Srujana Merugu, and Dhamendra Modha. A generalized maximum entropy approach to Bregman co-clustering and matrix approximation. *Journal of Machine Learning Research*, 2007.
- Andrew Barron and Thomas Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 1991.
- Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.
- Luc Devroye and Gábor Lugosi. *Combinatorial Methods in Density Estimation*. Springer, 2001.
- Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.
- Inderjit Dhillon, Subramanyam Mallela, and Dhamendra Modha. Information-theoretic co-clustering. In *ACM SIGKDD*, 2003.
- Peter Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.
- Michael Kearns, Yishay Mansour, Andrew Ng, and Dana Ron. An experimental and theoretical comparison of model selection methods. *Machine Learning*, 1997.
- John Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 2005.
- Andreas Maurer. A note on the PAC-Bayesian theorem. www.arxiv.org, 2004.
- McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 1999.
- Liam Paninski. Estimation of entropy and mutual information. *Neural Computation*, 2003.
- Yevgeny Seldin and Naftali Tishby. Multi-classification by categorical features via clustering. In *ICML*, 2008.
- Noam Slonim and Naftali Tishby. Document clustering using word clusters via the information bottleneck method. In *The 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2000.