Learning A Priori Constrained Weighted Majority Votes

Aurélien Bellet, Amaury Habrard, Emilie Morvant, Marc Sebban

Noname manuscript No.
(will be inserted by the editor)

Learning A Priori Constrained Weighted Majority Votes

Aurélien Bellet · Amaury Habrard · Emilie Morvant · Marc Sebban

Received: date / Accepted: date

Abstract Weighted majority votes allow one to combine the output of several classifiers or voters. MinCq is a recent algorithm for optimizing the weight of each voter based on the minimization of a theoretical bound over the risk of the vote with elegant PAC-Bayesian generalization guarantees. However, while it has demonstrated good performance when combining weak classifiers, MinCq cannot make use of the useful *a priori* knowledge that one may have when using a mixture of weak and strong voters. In this paper, we propose P-MinCq, an extension of MinCq that can incorporate such knowledge in the form of a constraint over the distribution of the weights, along with general proofs of convergence that stand in the sample compression setting for data-dependent voters. The approach is applied to a vote of k-NN classifiers with a specific modeling of the voters' performance. P-MinCq significantly outperforms the classic k-NN classifier, a symmetric NN and MinCq using the same voters. We show that it is also competitive with LMNN, a popular metric learning algorithm, and that combining both approaches further reduces the error.

Keywords Ensemble learning · Weighted majority vote · PAC-Bayesian bounds · Sample compression · Nearest neighbors

A. Bellet
Department of Computer Science
University of Southern California
E-mail: bellet@usc.edu

A. Habrard · M. Sebban
Université Jean Monnet de Saint-Etienne, Laboratoire Hubert Curien UMR CNRS 5516, 18 rue du Professeur Benoit Lauras, 42000 Saint-Etienne Cedex 2, France
E-mail: amaury.habrard@univ-st-etienne.fr
E-mail: marc.sebban@univ-st-etienne.fr

E. Morvant
Institute of Science and Technology (IST) Austria
Am Campus 1, 3400 Klosterneuburg, Austria
E-mail: emorvant@ist.ac.at

1 Introduction

A weighted majority vote is an ensemble method (Dietterich, 2000; Re and Valentini, 2012) where several classifiers (or *voters*) are assigned a specific weight. Such approaches are motivated by the idea that a careful combination can potentially compensate for the individual classifiers' errors and thus achieve better robustness and performance. For this reason, ensemble learning has been a prominent research area in machine learning and many methods have been proposed in the literature, among which Bagging (Breiman, 1996), Boosting (Schapire and Singer, 1999) or Random Forests (Breiman, 2001). The problem has also been studied from a Bayesian learning perspective, for instance with Bayesian model averaging (Haussler et al., 1994; Domingos, 2000). Multimedia analysis is an example of prolific application, for instance to combine classifiers learned from different modalities of the data (Atrey et al., 2010).

Even though combining weak classifiers such as in Boosting (Freund and Schapire, 1996) is supported by a solid theory, understanding when weighted majority votes perform better than a classic averaging of the voters is still a difficult question. In this context, PAC-Bayesian theory (McAllester, 1999) offers an appropriate framework to study majority votes and learn them in a principled way and with generalization guarantees. In particular, the recently-proposed MinCq (Laviolette et al., 2011) optimizes the weights of a set of voters $\mathcal{H}$ by minimizing a bound—the C-bound (Lacasse et al., 2007)—involving the first two statistical moments of the margin achieved on the training data. The authors show that minimizing this bound allows one to minimize the true risk of the weighted majority vote and boils down to a simple quadratic program. MinCq returns a *posterior* distribution on $\mathcal{H}$ that gives the weight of each voter. It is based on an *a priori* uniform belief on the relevance of the voters, which is well-suited when combining weak classifiers. For instance, it has been successfully applied to weighted majority votes of decision stumps and RBF kernel functions. However, this uniform prior is not appropriate when one wants to combine efficiently various classifiers with different levels of performance.

In this paper, we claim that MinCq can be extended to deal with variable-performing classifiers when one has an *a priori* belief on the voters. We generalize MinCq in two respects. First, we propose a new formulation by extending the original notion of aligned distribution (Germain et al., 2011) to $\mathbf{P}$-aligned distributions. $\mathbf{P}$ models a constraint over the distribution on the weights of the voters, allowing us to incorporate an *a priori* belief on each voter, constraining the posterior distribution. Our extension, called P-MinCq, does not induce any loss of generality and we show that this new problem can still be formulated in a efficient way as a quadratic program. Second, we extend the proofs of convergence of Laviolette et al. (2011) to the sample compression setting (Graepel et al., 2005), where the voters are built from training examples, such as NN classifiers. Our results use similar arguments as those proposed in (Germain et al., 2011; Laviolette and Marchand, 2007) but our setting requires a specific proof, since the results of Germain et al. (2011) are only valid for surrogate losses bounding the $0-1$ loss.

The second part of the paper makes use of these two general contributions to optimize a weighted majority vote over a set of k-NN classifiers ($k=\{1,2,\dots\}$) to hightlight the benefit of an *a priori* on the voters. We propose a suitable *a priori* constraint $\mathbf{P}$ modeling the fact that we have more confidence in close neighborhoods. The idea is to *a priori* constrain larger (resp. smaller) weights on classifiers with small (resp. large) values of k to reflect the belief that local neighborhoods convey more relevant information than distant ones, which cannot be modeled by the uniform belief used in MinCq. Using P-MinCq in this context constitutes an original approach to learning a robust combination of NN classifiers that achieves bet-

ter accuracy. This is confirmed by experiments conducted on twenty benchmark datasets: P-MinCq clearly outperforms k-NN, a symmetric version of it (Nock et al., 2003), as well as MinCq based on the same voters. Moreover, for high-dimensional problems, P-MinCq turns out to be quite robust to overfitting. We also show that it is competitive with the metric learning algorithm LMNN (Weinberger and Saul, 2009) and that plugging the learned distance into P-MinCq can further improve the results. Finally, we apply our approach to an object categorization dataset, on which P-MinCq again achieves good performance.

This paper is organized as follows. Section 2 reviews MinCq and its theoretical basis. In Section 3, we introduce P-MinCq, our extension of MinCq to $\mathbf{P}$-aligned distributions. We derive generalization bounds for the sample compression case in Section 4. Section 5 shows that MinCq does not perform well when using NN-based voters and presents a $\mathbf{P}$-aligned distribution that is suitable to this context. Experiments are presented in Section 6.

2 Notations and Background

2.1 Preliminaries

Throughout this paper, we consider the framework of the algorithm MinCq (Laviolette et al., 2011) for learning a weighted majority vote over a set of real-valued voters for binary classification problems. Let $\mathcal{X} \in \mathbb{R}^d$ be the *input space* of dimension d and $\mathcal{Y} = \{-1, +1\}$ be the *output space* (*i.e.,* the set of possible labels). S denotes the training sample made of m labeled examples $(\mathbf{x}, y)$ drawn *i.i.d* over $\mathcal{X} \times \mathcal{Y}$ according to a fixed and unknown distribution D. The distribution of S of size m is denoted by D^m. MinCq takes its roots from the PAC-Bayesian theory (first introduced by McAllester (1999)). Given a set of voters $\mathcal{H}$, this theory is based on a *prior distribution* P and a *posterior distribution* Q, both of support $\mathcal{H}$. P models the *a priori* information on the relevance of the voters: those that are believed to perform best have larger weights in P.[1] By taking into account the information carried by S, the learner aims at adapting P to get the posterior distribution Q that implies the Q-**weighted majority vote** with the best generalization performance.

Definition 1 Let $\mathcal{H} = \{h_1, \ldots, h_n\}$ be a set of voters (or classifiers) from $\mathcal{X}$ to $\mathbb{R}$. Let Q be a distribution over $\mathcal{H}$. A Q-**weighted majority vote** classifierfootnoteSometimes B_Q is called the Bayes classifier. B_Q is defined:

$$\forall \mathbf{x} \in \mathcal{X},\ B_Q(\mathbf{x}) = \text{sign}\left[\mathop{\mathbf{E}}_{h\sim Q} h(\mathbf{x})\right] = \text{sign}\left[\sum_{h\in\mathcal{H}} Q(h)h(\mathbf{x})\right].$$

The **true risk** $R_D(B_Q)$ over the pairs $(\mathbf{x}, y)$ *i.i.d.* according to D is:

$$R_D(B_Q) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} I[B_Q(\mathbf{x}) \neq y],$$

where $I[.]$ is an indicator function.

Laviolette et al. (2011) and Lacasse et al. (2007) make the link between the risk $R_D(B_Q)$ and the following notion of Q-margin which models the confidence of B_Q in its labeling.

[1] As we will see, a key limitation of MinCq is that it requires an *a priori* uniform belief on the weights.

Definition 2 (Laviolette et al., 2011) The Q-**margin** of an example $(\mathbf{x}, y)$ over Q is:

$$\mathcal{M}_Q(\mathbf{x}, y) = y \mathop{\mathbf{E}}_{h\sim Q} h(\mathbf{x}).$$

The first and second moments of the Q-margin are:

$$\mathcal{M}_Q^D = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} \mathcal{M}_Q(\mathbf{x}, y) = \mathop{\mathbf{E}}_{h\sim Q} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} yh(\mathbf{x}), \text{ and}$$
$$\mathcal{M}_{Q^2}^D = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} (\mathcal{M}_Q(\mathbf{x}, y))^2 = \mathop{\mathbf{E}}_{(h,h')\sim Q^2} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} h(\mathbf{x})h'(\mathbf{x}).$$

It is easy to see that B_Q correctly classifies an example $\mathbf{x}$ if the Q-margin is strictly positive. Thus, under the convention that if $\mathcal{M}_Q(\mathbf{x}, y) = 0$, then B_Q errs on $(\mathbf{x}, y)$, we get:

$$R_D(B_Q) = \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D} \left(\mathcal{M}_Q(\mathbf{x}, y) \leq 0\right). \tag{1}$$

Let us finally introduce the following necessary notations:

$$\mathcal{M}_h^D = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} yh(\mathbf{x}), \text{ and } \mathcal{M}_{h,h'}^D = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} h(\mathbf{x})h'(\mathbf{x}). \tag{2}$$

If we use the training sample $S \sim D^m$ instead of the unknown distribution D, we get the **empirical risk** $R_S(B_Q)$, the **empirical first and second moments of the** Q-**margin** $\mathcal{M}_Q^S$ and $\mathcal{M}_{Q^2}^S$, and the associated $\mathcal{M}_h^S$ and $\mathcal{M}_{h,h'}^S$.

2.2 MinCq and Theoretical Results

We now review three recent results of Laviolette et al. (2011); Lacasse et al. (2007), which constitute the building blocks of our contributions. The first one takes the form of a bound—the C-bound (Theorem 1)—over $R_D(B_Q)$. It shows that the true risk can be minimized by only considering the first two moments of the Q-margin. Then, following some PAC-Bayesian generalization bounds, Theorem 2 justifies that the posterior distribution Q can be learned by minimizing the empirical C-bound. Finally, the authors show that learning an optimal Q-weighted majority vote boils down to a simple quadratic program called MinCq.

The C-bound is obtained by making use of Equation (1) and the Cantelli-Chebychev's inequality (Devroye et al., 1996) applied on the random variable $\mathcal{M}_Q(\mathbf{x}, y)$.

Theorem 1 (The C-bound (Laviolette et al., 2011)) *For any distributions Q over a class $\mathcal{H}$ of functions and D over $\mathcal{X}\times\mathcal{Y}$, if $\mathcal{M}_Q^D > 0$ then $R_D(B_Q) \leq C_Q^D$ where:*

$$C_Q^D = \frac{\mathbf{Var}_{(\mathbf{x},y)\sim D}\left(\mathcal{M}_Q(\mathbf{x}, y)\right)}{\mathbf{E}_{(\mathbf{x},y)\sim D}\left(\mathcal{M}_Q(\mathbf{x}, y)\right)^2} = 1 - \frac{\left(\mathcal{M}_Q^D\right)^2}{\mathcal{M}_{Q^2}^D}.$$

$C_Q^S = 1 - \frac{\left(\mathcal{M}_Q^S\right)^2}{\mathcal{M}_{Q^2}^S}$ *is its empirical counterpart.*

Thus, minimizing the C-bound appears to be a nice strategy for learning a Q that implies a Q-weighted majority vote B_Q with low true risk. To justify this strategy, the authors derive a PAC-Bayesian generalization bound for C_Q^D. To do so, they assume a **quasi-uniform distribution** Q over an **auto-complemented** set of $2n$ voters $\mathcal{H} = \{h_1, \ldots, h_n, h_{n+1}, \ldots, h_{2n}\}$, where: $h_{k+n} = -h_k$ (auto-complementation) and $Q(h_k) + Q(h_{k+n}) = \frac{1}{n}$ (quasi-uniformity)

for every $k \in \{1, \ldots, n\}$. Note that, for the sake of simplicity, we will denote $Q(h_k)$ by Q_k. They claim that this assumption is not too strong a restriction and characterizes situations where, in the absence of ground truth, one gives **the same** *a priori* **belief** on the voters. Moreover, such distributions have two advantages. On the one hand, they allow us to get rid of the classic term which captures the complexity of $\mathcal{H}$.[2] This is a clear advantage since such a term can be a bad regularization (Laviolette et al., 2011). On the other hand, this assumption plays the role of a regularization by giving the same *a priori* belief on the voters and provides a simple way to avoid overfitting.

The generalization bound is then obtained by taking the lower (resp. upper) bound on $\mathcal{M}_Q^D$ together with the upper (resp. lower) bound on $\mathcal{M}_{Q^2}^D$ from the following theorem.

Theorem 2 (Laviolette et al. (2011)) *For any distribution D over $\mathcal{X} \times \mathcal{Y}$, any $m \geq 8$, any auto-complemented family $\mathcal{H}$ of B-bounded real-valued voters, for all quasi-uniform distribution Q on $\mathcal{H}$, and for any $\delta \in (0, 1]$, we have:*

$$\underset{S\sim D^m}{\mathbf{Pr}}\left(\left|\mathcal{M}_Q^D - \mathcal{M}_Q^S\right| \leq \frac{2B\sqrt{\ln\frac{2\sqrt{m}}{\delta}}}{\sqrt{2m}}\right) \geq 1-\delta,$$

$$\textit{and}\ \underset{S\sim D^m}{\mathbf{Pr}}\left(\left|\mathcal{M}_{Q^2}^D - \mathcal{M}_{Q^2}^S\right| \leq \frac{2B^2\sqrt{\ln\frac{2\sqrt{m}}{\delta}}}{\sqrt{2m}}\right) \geq 1-\delta.$$

The authors have proved that their setting does not induce any lack of generality. From Theorems 1 and 2, they suggest the minimization of the empirical C-bound under the constraint $\mathcal{M}_Q^S \geq \mu$. Due to the quasi-uniformity assumption, they show that this minimization problem is equivalent to solving a simple quadratic program involving only the first n voters of $\mathcal{H}$. Their algorithm MinCq is given in Algorithm 1. It consists in minimizing the denominator $\mathcal{M}_{Q^2}^S$, *i.e.*, the second moment of the Q-margin (Line 3), under the constraints $\mathcal{M}_Q^S = \mu$ (Line 4) and Q is quasi-uniform (Line 5). This leads to minimizing the C-bound and thus the true risk of the majority vote by only taking into account the diversity between the voters expressed by the empirical second moment.
The Q-weighted majority vote learned by MinCq is:

$$B_Q(\mathbf{x}) = \text{sign}\left[\sum_{k=1}^{n}\left(2Q_k - \frac{1}{n}\right)h_k(\mathbf{x})\right].$$

3 Generalization of MinCq to P-Aligned Distributions

Rather than constraining Q to be a quasi-uniform on the auto-complemented set of $2n$ voters $\mathcal{H}$ ($\forall k \in \{1, \ldots, n\},\ Q_k + Q_{k+n} = \frac{1}{n}$) as done in MinCq, we generalize this approach to any $\mathbf{P}$-aligned distribution Q: $\forall k \in \{1, \ldots, n\}$, $Q_k + Q_{k+n} = P_k$, where $\mathbf{P} = (P_1, \ldots, P_n)^\top$ sums to 1. In this context, $\mathbf{P}$ plays the role of an *a priori* belief on the voters.

[2] In the PAC-Bayesian theory, this term is related to the Kullback-Leibler divergence between the posterior distribution Q and the prior distribution P. See (Laviolette et al., 2011) for more details.

Algorithm 1 MinCq: a quadratic program for learning Q-weighted majority vote, under quasi-uniformity constraint

input A sample $S \sim D^m$, the first n voters of an auto-complemented set $\mathcal{H}$, a desired margin $\mu > 0$
output A posterior vector $\mathbf{Q} = (Q_1, \ldots, Q_n)^\top$.

$$\textbf{Solve} \quad \operatorname*{argmin}_{\mathbf{Q}} \ \mathbf{Q}^\top \mathbf{M}_S \mathbf{Q} - \mathbf{A}_S^\top \mathbf{Q}, \tag{3}$$

$$\textbf{s.t.} \quad \mathbf{m}_S^\top \mathbf{Q} = \frac{\mu}{2} + \frac{1}{2n} \sum_{k=1}^{n} \mathcal{M}^S_{h_k}, \tag{4}$$

$$\forall k \in \{1, \ldots, n\}, \quad 0 \leq Q_k \leq 1/n, \tag{5}$$

where $\mathbf{Q} = (Q_1, \ldots, Q_n)^\top$ is the vector of the first n weights Q_k, $\mathbf{M}_S$ the $n \times n$ matrix formed by $\mathcal{M}^S_{h_k,h_{k'}}$ for $(k, k') \in \{1, \ldots, n\}^2$ (as defined in Equation (2)), $\mathbf{m}_S = \left(\mathcal{M}^S_{h_1}, \ldots, \mathcal{M}^S_{h_n}\right)^\top$, and:

$$\mathbf{A}_S = \left(\frac{1}{nm} \sum_{k=1}^{n} \mathcal{M}^S_{h_1,h_k}, \ldots, \frac{1}{nm} \sum_{k=1}^{n} \mathcal{M}^S_{h_n,h_k} \right)^\top .$$

3.1 Expressiveness of $\mathbf{P}$-aligned distributions

We generalize the setting of Laviolette et al. (2011) for quasi-uniform distributions to any $\mathbf{P}$-aligned distribution on a set of auto-complemented classifiers $\mathcal{H}$, in fact this constraint does not restrict the possible outcomes of an algorithm that would minimize C^S_Q.

Proposition 1 *For all distributions Q on $\mathcal{H}$, there exists a $\mathbf{P}$-aligned distribution Q' on the auto-complemented $\mathcal{H}$ that provides the same majority vote as Q, and that has the same empirical and true C-bound values.*

Proof It follows from Proposition 4 of (Germain et al., 2011) and is given in Appendix A.2.

From this proposition, similarly as for MinCq, it is then justified that under the constraint $\mathcal{M}^S_Q = \mu$, the C-bound can be optimized by minimizing the second moment $\mathcal{M}^S_{Q^2}$ of the Q-margin. This is done by solving the quadratic program P-MinCq described in the following.

3.2 The quadratic program P-MinCq

P-MinCq is described in Algorithm 2. Similarly to MinCq, thanks to the $\mathbf{P}$-aligned assumption, we only need to cope with the first n voters in $\mathcal{H}$. The objective function (Line (6)) minimizes the second moment of the Q-margin while the first constraint (Line (7)) enforces a margin equal to μ. Note that the left-hand side of this constraint is the weighted average (with weights of $2Q_k - P_k$) of the individual margins ($\mathcal{M}_{h_k}$). Finally, Line (8) restricts Q to be $\mathbf{P}$-aligned. The proof of derivation of the algorithm can be found in Appendix A.3.
The Q-weighted majority vote learned by P-MinCq is:

$$B_Q(\mathbf{x}) = \operatorname{sign}\left[\sum_{k=1}^{n} (2Q_k - P_k)\, h_k(\mathbf{x}) \right].$$

The next section addresses the generalization guarantees for P-MinCq.

Algorithm 2 P-MinCq: quadratic program for learning Q-weighted majority vote, under $\mathbf{P}$-aligned constraint

input A sample $S \sim D^m$, the first n voters of an auto-complemented set $\mathcal{H}$, a desired margin $\mu > 0$, a prior vector $\mathbf{P} = (P_1, \ldots, P_n)^\top$, a matrix M_S of size $n \times n$ made of elements $\mathcal{M}^S_{h_k, h_{k'}}$.

output A posterior vector $\mathbf{Q} = (Q_1, \ldots, Q_n)^\top$.

$$\textbf{Solve} \quad \underset{\mathbf{Q}}{\operatorname{argmin}} \ (\mathbf{Q} - \mathbf{P})^\top \mathbf{M}_S \mathbf{Q}, \tag{6}$$

$$\textbf{s.t.} \quad \mathbf{m}_S^\top (2\mathbf{Q} - \mathbf{P}) = \mu, \tag{7}$$

$$\forall k \in \{1, \ldots, n\}, \ 0 \leq Q_k \leq P_k, \tag{8}$$

where $\mathbf{m}_S^\top = (\mathcal{M}_{h_1}, \ldots, \mathcal{M}_{h_n})^\top$.

4 PAC-Bayesian Generalization Guarantees under Sample Compression

The proof of the generalization bounds of Theorem 2 is still valid for $\mathbf{P}$-aligned distribution Q over data-independent voters. Indeed, it only makes use of the $\mathbf{P}$-aligned assumption corresponding to $Q_k + Q_{k+n} = P_k + P_{k+n}$.[3] This theorem is nevertheless not valid in the sample compression setting, where the set of voters is data-dependent (such as the set of k-NN classifiers). Laviolette et al. (2011) have argued that it can be extended to this setting by using techniques from (Laviolette and Marchand, 2007). This section is devoted to derive generalization bounds for P-MinCq in this sample compression setting, allowing us to deal with data-dependent voters. Our result is rather general (and not restricted to k-NN voters). It differs from previous PAC-Bayesian results with sample compressed classifiers (Graepel et al., 2005; Laviolette and Marchand, 2007; Germain et al., 2011) in that it is tailored to the first two moments of the $\mathbf{Q}$-margin with $\mathbf{P}$-aligned distributions.

4.1 Sample compression setting

In the *sample compression framework* (Floyd and Warmuth, 1995) the learning algorithm $\mathcal{A}$ has access to a data-dependent set of classifiers. Each classifier is then represented by two elements: a **compression sequence** which is a sequence of examples, and a **message** representing the additional information needed to obtain the classifier from the compression sequence. Then, we can define a **reconstruction function** able to output a classifier from a compression sequence and a message. More formally, a learning algorithm $\mathcal{A}$ is called a **compression scheme** if it is defined as follows.

Definition 3 Let $S \in (\mathcal{X} \times \mathcal{Y})^m = \mathcal{Z}^m$ be the learning sample of size m. We define $\mathbf{J}_m$ to be the set containing all the possible vectors of indices:

$$\mathbf{J}_m = \bigcup_{i=1}^{m} \left\{ (j_1, \ldots, j_i) \in \{1, \ldots, m\}^i \right\}.$$

Given a family of hypothesis $\mathcal{H}^{\mathcal{S}}$ from $\mathcal{X}$ to $\mathcal{Y}$ and an index vector $\mathbf{j} \in \mathbf{J}_m$, let $S_{\mathbf{j}}$ be the subsequence indexed by $\mathbf{j}$, $S_{\mathbf{j}}$ is called the **compression sequence**:

$$S_{\mathbf{j}} = (\mathbf{z}_{j_1}, \ldots, \mathbf{z}_{j_i}).$$

[3] See (Laviolette et al., 2011) for more details.

An algorithm $\mathcal{A}: \mathcal{Z}^{(\infty)} \mapsto \mathcal{H}^{\mathcal{S}}$ is a **compression scheme** if, and only if, there exists a triplet $(\mathcal{C}, \mathcal{R}, \omega)$ such that for all training sample S:

$$\mathcal{A}(S) = \mathcal{R}\left(S_{\mathcal{C}(S)}, \omega\right),$$

where $\mathcal{C} : \mathcal{Z}^{(\infty)} \mapsto \bigcup_{m=1}^{\infty} \mathbf{J}_m$ is the **compression function**, $\mathcal{R} : \mathcal{Z}^{(\infty)} \times \Omega_{S_{\mathcal{C}(S)}} \mapsto \mathcal{H}^{\mathcal{S}}$ the **reconstruction function**, and ω is a **message** chosen from the set $\Omega_{S_{\mathcal{C}(S)}}$ (*a priori* defined) of all messages that can be supplied with the compression sequence $S_{\mathcal{C}(S)}$.

Put into words, given a learning sample $S \sim D^m$, a sample compression scheme is a reconstruction function $\mathcal{R}$ mapping a compression sequence $\mathcal{C}(S) = S_{\mathbf{j}}$ to some set $\mathcal{H}^S$ of functions $h^{\omega}_{S_{\mathbf{j}}}$ such that $\mathcal{A}(S) = \mathcal{R}\left(S_{\mathbf{j}}, \omega\right) = h^{\omega}_{S_{\mathbf{j}}}$. For example, k-NN classifiers can be reconstructed from a compression sequence only, which encodes the nearest neighbors (Floyd and Warmuth, 1995; Graepel et al., 2005). Other classifiers, such as the decision list machines (Marchand and Sokolova, 2005), need both a compression sequence and a message string. In the next section, we consider the general setting to avoid any loss of generality.

4.2 PAC-Bayesian generalization bounds under sample compression

Let $S_{\mathbf{j}}$ be a sample compression sequence consisting of $|\mathbf{j}|$ elements of the learning sample S. In the *PAC-Bayesian sample compression setting*, the risks R_D and R_S can be biased by these elements: we often prefer to compute the empirical risk R_S from $S \backslash S_{\mathbf{j}}$ (Laviolette and Marchand, 2007). However, in order to derive risk bounds in such a situation, Germain et al. (2011) have proposed another strategy by directly considering the bias. As mentioned in the introduction, we cannot apply their result to our setting. Indeed, it is valid for loss functions defining a surrogate of the $0-1$ loss, which is not suited for the second moment of the margin we have to consider. Moreover, it depends on the value of the surrogate at -1, which may lead to a degenerate bound (this does not occur in our bounds).

The derivation of our result is nevertheless based on a similar setting: given a sample S, we consider $\mathcal{H}^S$ the set of all possible classifiers $h^{\omega}_{S_{\mathbf{j}}} = \mathcal{R}(S_{\mathbf{j}}, \omega)$ such that $\omega \in \Omega_{S_{\mathbf{j}}}$. We denote by $Q_{\mathbf{J}_m}(\mathbf{j})$, the probability that a compression sequence $S_{\mathbf{j}}$ is chosen by Q, and $Q_{S_{\mathbf{j}}}(\omega)$ the probability of choosing the message ω given $S_{\mathbf{j}}$. Then, we have:

$$Q_{\mathbf{J}_m}(\mathbf{j}) = \int_{\omega \in \Omega_{S_{\mathbf{j}}}} Q(h^{\omega}_{S_{\mathbf{j}}}) d\omega, \quad \text{and} \quad Q_{S_{\mathbf{j}}}(\omega) = Q(h^{\omega}_{S_{\mathbf{j}}} | S_{\mathbf{j}}).$$

In the usual PAC-Bayesian setting, the risk bounds depend on the prior distribution P over the set $\mathcal{H}^S$. This prior distribution is supposed to be known before observing the learning sample S, implying P independent from S. However, in our setting the classifiers in $\mathcal{H}^S$ are data-dependent. To tackle this problem, we propose to follow the principle of Laviolette and Marchand (2007); Germain et al. (2011) by considering a prior distribution defined by a pair: $\left(P_{\mathbf{J}_m}, (P_{S_{\mathbf{j}}})_{\mathbf{j} \in \mathbf{J}_m}\right)$, where $P_{\mathbf{J}_m}$ is a distribution over $\mathbf{J}_m$, and for all possible compression sequence $S_{\mathbf{j}}$, $P_{S_{\mathbf{j}}}$ is a distribution over $\Omega_{S_{\mathbf{j}}}$. Given a training sample S, the data-independent prior distribution P corresponds to the distribution on $\mathcal{H}^S$ associated with the prior $\left(P_{\mathbf{J}_m}, (P_{S_{\mathbf{j}}})_{\mathbf{j} \in \mathbf{J}_m}\right)$, then we have: $P(h^{\omega}_{S_{\mathbf{j}}}) = P_{\mathbf{J}_m} P_{S_{\mathbf{j}}}(\omega)$.

Definition 4 In the sample compression setting, the Q-**margin** of a point $(\mathbf{x}, y)$ over Q is:

$$\mathcal{M}_Q(\mathbf{x}, y) = y \underset{h^{\omega}_{S_{\mathbf{j}}} \sim Q}{\mathbf{E}} h^{\omega}_{S_{\mathbf{j}}}(\mathbf{x}).$$

The first two moments $\mathcal{M}_Q^D$ and $\mathcal{M}_{Q^2}^D$ of the Q-margin are defined similarly as before:

$$\mathcal{M}_Q^D = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} \mathcal{M}_{\mathcal{Q}}(\mathbf{x},y) \text{ and } \mathcal{M}_{Q^2}^D = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} (\mathcal{M}_Q(\mathbf{x},y))^2.$$

In our setting, we assume **P**-aligned distributions on an auto-complemented set $\mathcal{H}^S$. For each classifier $h_S^\omega \in \mathcal{H}^S$, we denote its complement by $-h_S^\omega$. Given S, the associated message set is $\Omega_S \times \{+,-\}$ and $\forall \sigma \in \Omega_S$, $h_S^{(\sigma,+)} = -h_S^{(\sigma,-)}$. We now give the main result of this section.

Theorem 3 *For any distribution D over $\mathcal{X} \times \mathcal{Y}$, any $m \geq 8$, any auto-complemented set $\mathcal{H}^S$ of B-bounded real valued voters of sample compression size at most $|\mathbf{j}^{\max}| < \frac{m}{2}$, for all* **P***-aligned distribution Q on $\mathcal{H}^S$, and for any $\delta \in (0,1]$, we have:*

$$\mathop{\mathbf{Pr}}_{S\sim D^m}\left(\left|\mathcal{M}_Q^D - \mathcal{M}_Q^S\right| \leq \frac{2B\sqrt{\frac{|\mathbf{j}^{\max}|}{B} + \ln\left(\frac{2\sqrt{m}}{\delta}\right)}}{\sqrt{2(m-|\mathbf{j}^{\max}|)}}\right) \geq 1-\delta, \tag{9}$$

$$\mathop{\mathbf{Pr}}_{S\sim D^m}\left(\left|\mathcal{M}_{Q^2}^D - \mathcal{M}_{Q^2}^S\right| \leq \frac{2B^2\sqrt{\frac{2|\mathbf{j}^{\max}|}{B^2} + \ln\left(\frac{2\sqrt{m}}{\delta}\right)}}{\sqrt{2(m-2|\mathbf{j}^{\max}|)}}\right) \geq 1-\delta. \tag{10}$$

Proof Deferred to Appendix A.4.

For data-independent classifiers, *i.e.* $|\mathbf{j}^{\max}| = 0$, we recover Theorem 2. As expected, the theorem indicates that when the compression size $|\mathbf{j}^{\max}|$ is large, the bound becomes looser, suggesting that the compression size should not be too large to preserve consistency. Note that the bound B over the classifiers' output can generally be controlled by the use of appropriate normalization.

In the next section, we instantiate P-MinCq in the specific k-NN setting by introducing a rather intuitive but statistically well-founded *a priori* constraint $\mathbf{P}$.

5 Instantiation of P-MinCq for Nearest Neighbor Classifiers

5.1 Limitations of MinCq in the context of nearest neighbor classifiers

At first sight, one may think that MinCq is a good way to overcome two limitations of k-NN classifiers. First, while the theory tells us that the higher k, the better the convergence to the optimal bayesian risk, this holds only asymptotically. In practice the choice of k requires special care. Therefore, optimizing a Q-weighted majority vote, where the set of voters $\mathcal{H}$ consists of the k-NN classifiers ($k = \{1,2,\dots\}$), would prevent us from tuning k while offering a principled way to combine these classifiers.[4] Second, by making use of the PAC-Bayesian setting, the minimization of the C-bound provides generalization guarantees that cannot be obtained with a standard k-NN algorithm in finite-sample situations.

We conduct a preliminary experimental study to compare a standard k-NN classifier (where k is tuned by cross-validation) with MinCq (see Section 6 for details on the setup).

[4] Note that other strategies may be used to define the voters, *e.g.*, the n^{th} neighbor can be the n^{th} voter.

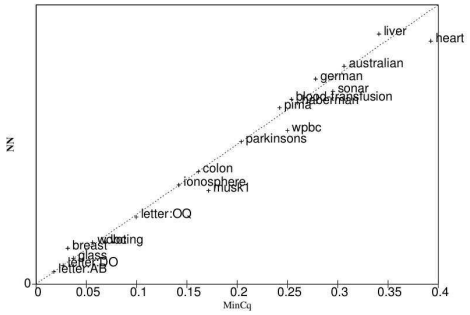

Fig. 1 Comparison of MinCq VS **NN**. Each point in the scatter plot shows the test error rate of the algorithms on a single dataset. A dataset above the bisecting line is in favor of MinCq

Over twenty datasets, MinCq achieves an average classification error of 18.18% against 17.88% for k-NN (see Table 1 for more details). It is worth noting that using a Student paired t-test, we cannot statistically distinguish between the two approaches. This is also confirmed by a sign test, which gives a record win/loss/tie equal to 7/6/7 leading to a p-value of about 0.5, as illustrated by Figure 1. This serie of experiments clearly shows that MinCq performs no better than a single well-tuned k-NN classifier.

We claim that these disappointing results can be explained by the fact that the quasi-uniformity assumption on Q is not appropriate to settings where one has an *a priori* belief on the relevance of the voters, which is typically the case in NN classification. Indeed, for obvious reasons, close neighborhoods are likely to provide more relevant information than distant ones. We propose to overcome these limitations by using an instantiation of P-MinCq based on a constraint $\mathbf{P}$ suitable for NN classification.

5.2 A statistically well-founded constraint $\mathbf{P}$

In standard k-NN classification, the theory tells us that the higher k, the better the convergence to the optimal bayesian risk. However, this property holds only asymptotically, *i.e.*, when the size m of the training sample goes to infinity. In practice, training data is limited and one has to set k carefully. On the one hand, we want to use a large value of k to obtain a reliable estimate. On the other hand, only points in a very close neighborhood lead to an accurate classification rule. Several theoretical and experimental studies in the literature have tried to analyze this trade-off between small and large values of k. As suggested by Duda et al. (2001), a good solution consists in using a small fraction of the training examples, equal to about $\sqrt{m/|\mathcal{Y}|}$ neighbors, where $|\mathcal{Y}|$ is the number of classes.

The context is slightly different in P-MinCq, since we aim at linearly combining k-NN classifiers ($k = 1, 2, \ldots$). Rather than setting k, we aim at choosing a suitable constraint $\mathbf{P}$, which plays the role of an *a priori* belief on the voters. As suggested by Devroye et al. (1996), in a weighted nearest neighbor rule, nearer neighbors should provide more information than distant ones. Following this, we propose the following constraint $\mathbf{P}$ (normalized so that they sum to 1):

$$\forall k \geq 1, \quad P_k = 1/k. \tag{11}$$

$\mathbf{P}$ concentrates the weights on voters that are based on a small fraction of the training data, *i.e.*, points in a close neighborhood (as suggested by Duda et al. (2001)), but also takes into

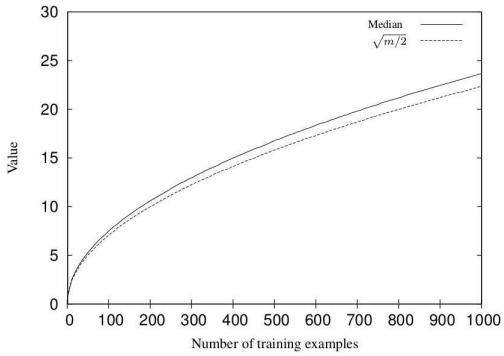

Fig. 2 Comparison between the median of the harmonic series $\sum_{x=1}^{m} \frac{1}{x}$ and $\sqrt{m/2}$

account (to a smaller extent) the information provided by (potentially) the entire training set. To justify this choice, we establish in the following a strong relationship between Equation (11) and the popular choice $\sqrt{m/2}$ for k in k-NN binary classification. Our analysis is based on the characterization of $\mathbf{P}$ by its median M, which corresponds to the number of neighbors involved in the voters accumulating half of the total weight. While defining the median of a continuous distribution is rather straightforward, finding it in the discrete case of interest (*i.e.*, where $x \in \{1, \ldots, m\}$) is slightly more tricky and requires an approximation. Let us define $H_M = \sum_{x=1}^{M} \frac{1}{x}$ and $H_m = \sum_{x=1}^{m} \frac{1}{x}$. They correspond to the sum of terms of a harmonic series for which no closed form is available. However, using the partial sums of the series, for all n we can define H_n such that: $H_n = \sum_{x=1}^{n} \frac{1}{x} = \ln(n) + \gamma + \epsilon_n,$ where γ is the Euler-Mascheroni constant ($\gamma \simeq 0.5772156$) and $\epsilon_n \sim \frac{1}{2n}$. Therefore, we have:

$$
\begin{aligned}
H_M = \frac{1}{2} H_m \Leftrightarrow & \sum_{x=1}^{M} \frac{1}{x} = \frac{1}{2} \sum_{x=1}^{m} \frac{1}{x} \\
\Leftrightarrow & \ln(M) + \gamma + \epsilon_M = \frac{1}{2}(\ln(m) + \gamma) + \frac{1}{2}\epsilon_m \\
\Leftrightarrow & \ln(M) = \ln(\sqrt{m}) - \frac{1}{2}\gamma + \frac{1}{2}\epsilon_m - \epsilon_M \\
\Rightarrow & \ln(M) \cong \ln(\sqrt{m}) - \frac{1}{2}\gamma + \frac{1}{4m} - \frac{1}{2M} \quad \text{(using } \epsilon_n \sim \tfrac{1}{2n}\text{)} \\
\Rightarrow & \ln(M) \leq \ln(\sqrt{m}) - \frac{1}{2}\gamma - \frac{1}{4m} \quad \text{(since Equation (11) } \Rightarrow M \leq m/2\text{)} \\
\Rightarrow & M \leq \sqrt{m \exp(-\gamma) \exp\left(-\frac{1}{4m}\right)} \simeq \sqrt{\frac{m}{2}}. \qquad (12)
\end{aligned}
$$

The main information provided by Equation (12) is that the approximation of the median of $\mathbf{P}$ is very close to $\sqrt{m/2}$, the value suggested for k in the k-NN rule for binary classification problems. Figure 2 shows a graphical illustration of the closeness between the median of the harmonic series and $\sqrt{m/2}$. We have thus established a strong relationship between a classic choice for k in standard k-NN classification and our $\mathbf{P}$ constraint in a weighted majority vote of k-NN voters. The next section will feature a large comparative experimental study that validates our choice for $\mathbf{P}$.

Before that, recall that the generalization bound derived in Section 4 suggests to limit the prototype set for the k-NN classifiers. A first approach could be to divide the learning

sample in two sets: one for defining the k-NN classifiers and one for learning the parameters of the model. However, this strategy does not stand in the sample compression scheme and has the disadvantage to discard useful information. Another solution is to apply—for each k-NN voter—some prototype selection or reduction techniques (Duda et al., 2001) in order to remove training examples that do not change the labeling of any test example. This implies that each k-NN must use its own compressed sample corresponding to a subset of the training sample S. However, in addition to its computational cost, this strategy is not always relevant in the context of NN since it may be difficult to obtain a good (*i.e.* small) compression scheme for some distributions. Nevertheless, in the particular setting we consider for k-NN, we have noticed that using large $|\mathbf{j}^{\max}|$ (even equals to m) does not influence the practical performance of P-MinCq.

6 Experimental Results

In this section, we propose a comparative study of P-MinCq applied to the context of NN classification (as described in Section 3). We compare it against four different approaches.

- The standard Nearest Neighbor algorithm (NN) which plays the role of the baseline.
- The Symmetric Nearest Neighbor algorithm (Nock et al., 2003) (SNN), a variant of NN where the class of an instance x is determined by the majority class among the training points that belong to the k-neighborhood of x (like in NN) plus those that include x in their own k-neighborhood.
- Large Margin Nearest Neighbor (Weinberger and Saul, 2009) (LMNN) which learns a Mahalanobis distance by optimizing the k-NN training error (with a safety margin). Then, k-NN is applied using the learned distance. Note that LMNN has been shown to be competitive with a RBF kernel SVM.
- MinCq (Laviolette et al., 2011) which considers a quasi-uniform distribution.

We evaluate these methods on twenty benchmark datasets and an object categorization task.

6.1 Benchmark datasets

Experimental setup. These twenty binary classification datasets are of varying domain and difficulty, mostly taken from the UCI Machine Learning Repository.[5] We compute neighborhoods using the standard Euclidean distance. We randomly split each dataset into 50% training and 50% test data, except for letterAB, letterDO and letterOQ for which we split $20\%/80\%$. We tune the following parameters by 10-fold cross-validation on the training set: the margin parameter μ for MinCq and P-MinCq (among 14 values between $.0001$ and $.5$) and the parameter k for k-NN and LMNN (among $\{1, \ldots, 10\}$). The trade-off parameter of LMNN was set to $.5$, as done by Weinberger and Saul (2009).

Results. We report the results in Table 1. We make the following remarks. First, P-MinCq significantly outperforms a standard NN classifier. On average over the datasets, P-MinCq achieves a classification error of 16.89% while NN reaches a level of 17.88%. Using a Student paired t-test, this difference is statistically significant with a p-value of $.06$. This is further supported by a sign test, which gives a record win/loss/tie equals to 12/5/3 leading to

[5] `http://archive.ics.uci.edu/ml/`

Table 1 Error rates of NN, SNN, LMNN, MinCq and P-MinCq on twenty datasets

Dataset	NN	SNN	LMNN	MinCq	P-MinCq
australian	.3121	.3324	.2746	.3064	.2919
blood	.2647	.2487	.2674	.2540	.2567
breast	.0514	.0200	.0400	.0314	.0257
colon	.1613	.1290	.2258	.1613	.1290
german	.2940	.3040	.2760	.2780	.2720
glass	.0370	.0648	.0648	.0370	.0370
haberman	.2597	.2532	.2922	.2597	.2727
heart	.3481	.3926	.2148	.3926	.3556
ionosphere	.1420	.1591	.1193	.1420	.0795
letter:AB	.0176	.0143	.0151	.0176	.0176
letter:DO	.0268	.0293	.0126	.0268	.0260
letter:OQ	.0961	.0961	.0334	.0995	.0892
liver	.3584	.3468	.3584	.3410	.3584
musk1	.1339	.1464	.2092	.1715	.1297
parkinsons	.2041	.2143	.1531	.2041	.2347
pima	.2526	.2474	.2604	.2422	.2370
sonar	.2762	.2952	.0762	.2952	.2000
voting	.0596	.0596	.0413	.0688	.0688
wdbc	.0596	.0842	.0491	.0561	.0456
wpbc	.2200	.2500	.2300	.2500	.2500
Avg. error	.1788	.1844	.1607	.1818	.1689
Avg. rank	2.9	3.1	2.65	2.9	2.25

a p-value of $.07$. P-MinCq also outperforms SNN despite the fact that the latter performs well on a few datasets (p-value of $.01$ with a Student test and $.24$ with a sign test). Furthermore, P-MinCq performs significantly better than MinCq with a p-value of $.02$ using a Student test. With a sign test, the p-value is about $.03$ with a record win/loss/tie equals to 12/4/4. This shows the usefulness of our generalization of MinCq to $\mathbf{P}$-aligned distributions, and that $P_i = \frac{1}{i}$ is a suitable *a priori* distribution in the context of NN. Finally, despite the fact that P-MinCq is not a metric learning algorithm, it is competitive with LMNN ($.1689$ versus $.1607$ with a p-value of about $.10$ with a Student test). A sign test leads to a p-value of $.5$, indicating that one method is equally likely to perform better than the other.

In fact, we claim that P-MinCq and LMNN are rather complementary. Indeed, on the one hand, LMNN is a metric learning algorithm that can tweak the neighborhoods of the points (sometimes with great success, *e.g.*, heart, parkinsons or sonar) but may perform worse than NN, especially because it often overfits when dimensionality is high (*e.g.*, colon or musk1). On the other hand, P-MinCq does not change the neighborhoods of the points but combines several nearest neighbor rules, and as a combination of classifiers, appears to be quite stable (as shown at the bottom of Table 1, it achieves the best average rank) and robust to overfitting. To highlight how P-MinCq and LMNN complement each other, we perform an additional series of experiments aiming at combining LMNN and P-MinCq when this seems relevant. To do so, we make use of the validation performance: if LMNN performs better than P-MinCq, then we plug the distance learned by LMNN in P-MinCq (otherwise we keep the standard Euclidean distance). We report the results in Table 2. The combination LMNN+P-MinCq outperforms all other methods, including LMNN alone (p-values of $.05$ with a Student test and $.17$ with a sign test). Notice that on some datasets where LMNN was by far the best performing method in the first series of experiments (*e.g.*, on heart, parkinsons or voting), LMNN+P-MinCq is able to further improve these results.

Table 2 Error rates of LMNN and LMNN+P-MinCq on twenty datasets

Dataset	LMNN	LMNN+P-MinCq
australian	**.2746**	.2832
blood	**.2674**	.2701
breast	.0400	**.0257**
colon	.2258	.2258
german	**.2760**	.2820
glass	.0648	**.0370**
haberman	.2922	**.2727**
heart	.2148	**.1926**
ionosphere	.1193	**.0795**
letter:AB	.0151	.0151
letter:DO	.0126	**.0084**
letter:OQ	**.0334**	.0386
liver	.3584	.3584
musk1	.2092	**.1297**
parkinsons	.1531	**.1020**
pima	.2604	**.2370**
sonar	**.0762**	.0952
voting	.0413	**.0367**
wdbc	.0491	**.0456**
wpbc	**.2300**	.2800
Avg. error	.1607	**.1508**

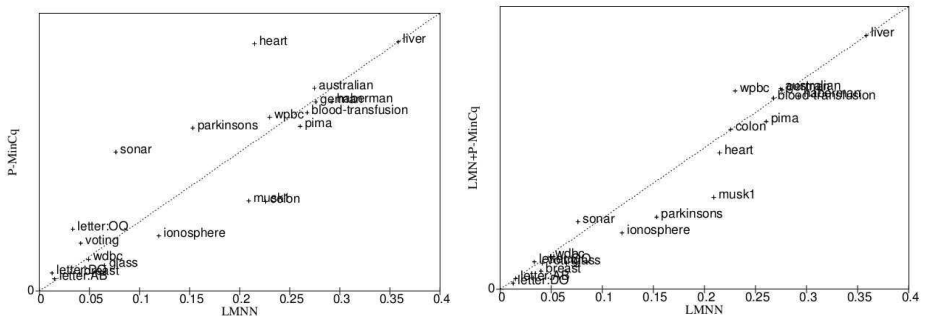

Fig. 3 Comparison of P-MinCq versus LMNN (left) and P-MinCq+LMNN versus LMNN (right)

6.2 Object categorization

Experimental setup. We provide additional experiments on Graz-01 (Opelt et al., 2004), a popular object categorization database that has two object-class (bike and person) and a background class. It is known to have large intra-class variation and significant background clutter (see Figure 4). The tasks are bike *vs* non-bike and person *vs* non-person and we follow experimental setup from (Opelt et al., 2004): for each object, we randomly sample 100 positive examples and 100 negative examples (of which 50 are drawn from the other object and 50 from the background). Images are represented as frequency histograms of 200 visual words built from SIFT interest points. We thus compute neighborhoods using two popular histogram distances: the χ^2 and the intersection distances.

Results. We report the results in Table 3, averaged over 10 runs. P-MinCq is again the most stable method and also the best on average across tasks and distance measures. Indeed, it

Fig. 4 Some examples of bikes (left column), persons (middle) and background (right) taken from Graz-01. Only parts of the objects of interest may be visible, and the background class features difficult counter-examples to the bike class, such as motorbikes

Table 3 Error rates of NN, SNN, MinCq and P-MinCq on the Graz-01 database, averaged over 10 runs.

Distance	Task	NN	SNN	MinCq	P-MinCq
χ^2	bike	.2310	.2090	.2160	.2095
χ^2	person	.2385	.2305	.2730	.2250
Intersection	bike	.2260	.2185	.2130	.2055
Intersection	person	.2350	.2370	.3180	.2255
Avg. error		.2326	.2238	.2550	.2164

significantly outperforms MinCq (p-value smaller than $.01$ with a Student test), again illustrating the importance of a good prior $\mathbf{P}$ for learning the majority vote. Moreover, P-MinCq performs significantly better than NN (p-value smaller than $.01$ with a Student test) and to a smaller extent than SNN (p-value of $.13$). It is worth noting that SNN performs rather well on this database: with large intra-class variation, it seems that extending the neighborhood can pay off. However, while the symmetry heuristic used by SNN is not relevant for all datasets, P-MinCq provides a principled and robust alternative.

7 Conclusion and Future Research

In this work, we have proposed a novel approach called P-MinCq for learning a weighted majority vote over variable-performing classifiers in the context of a recent algorithm MinCq which finds its grounds in the PAC-Bayesian theory. Our method is based on a generalization of MinCq to $\mathbf{P}$-aligned distributions allowing us to incorporate an *a priori* knowledge in the form of a distribution on the voters. This approach does not restrict the expressiveness of the majority vote and we have provided generalization guarantees for data-dependent voters such as k-NN classifiers. Moreover, we have defined a specific $\mathbf{P}$-aligned distribution adapted to the case of k-NN and provided experimental evidence of its good behavior.

Many promising perspectives arise from this work. First, the setting proposed in this paper is general enough to be used to combine virtually any set of classifiers (provided that they are bounded). For instance, our approach allows one to combine strong and weak classifiers and incorporate some *a priori* knowledge about their performance. Another interesting application is multi-view learning (Xu et al., 2013; Sun, 2013), where P-MinCq could be used to combine classifiers (such as SVM) trained on multi-modal data coming from different

sources and/or feature types (Morvant et al., 2014). In this case, $\mathbf{P}$ could encode the prior knowledge about the relative relevance of each modality for the task at hand. In general, in the absence of background knowledge, we note that defining a relevant $\mathbf{P}$ distribution for a set of learners can be difficult. Developing strategies to automatically assess $\mathbf{P}$ from (held-out) data could be very helpful in practice (Lever et al., 2013).

It would also be interesting to combine P-MinCq with other metric learning algorithms, such as the recent χ^2 distance learning method for histogram data (Kedem et al., 2012). Lastly, extending P-MinCq to a multi-class setting is also of high interest. However, this requires margin and loss definitions tailored to multi-class problem that imply technical difficulties, with the need of different theoretical tools such as in (Morvant et al., 2012).

Acknowledgements

This work was in parts funded by the European Research Concil under the European Unions Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no 308036, and by the french project SoLSTiCe ANR-13-BS02-01 of the ANR. Most of the work in this paper was carried out while Aurélien Bellet was affiliated with Université Jean Monnet de Saint-Etienne, Laboratoire Hubert Curien UMR CNRS 5516, France.

References

Atrey PK, Hossain MA, El Saddik A, Kankanhalli MS (2010) Multimodal Fusion for Multimedia Analysis: a Survey. Multimedia systems 16(6):345–379

Breiman L (1996) Bagging predictors. Machine Learning 24(2):123–140

Breiman L (2001) Random forests. Machine Learning 45(1):5–32

Devroye L, Györfi L, Lugosi G (1996) A Probabilistic Theory of Pattern Recognition. Springer-Verlag

Dietterich TG (2000) Ensemble methods in machine learning. In: Multiple Classifier Systems, pp 1–15

Domingos P (2000) Bayesian averaging of classifiers and the overfitting problem. In: International Conference on Machine Learning, p 223230

Duda R, Hart P, Stork D (2001) Pattern classification. Pattern Classification and Scene Analysis: Pattern Classification, Wiley, URL books.google.fr/books?id=YoxQAAAAMAAJ

Floyd S, Warmuth MK (1995) Sample Compression, Learnability, and the Vapnik-Chervonenkis Dimension. Machine Learning 21(3):269–304

Freund Y, Schapire R (1996) Experiments with a new boosting algorithm. In: International Conference on Machine Learning, pp 148–156

Germain P, Lacoste A, Laviolette F, Marchand M, Shanian S (2011) A PAC-Bayes Sample Compression Approach to Kernel Methods. In: International Conference on Machine Learning

Graepel T, Herbrich R, Shawe-Taylor J (2005) PAC-Bayesian Compression Bounds on the Prediction Error of Learning Algorithms for Classification. Machine Learning 59(1-2):55–76

Haussler D, Kearns M, Schapire R (1994) Bounds on the sample complexity of bayesian learning using information theory and the vc dimension. Machine Learning 14(1):83–113

Kedem D, Tyree S, Weinberger K, Sha F, Lanckriet G (2012) Non-linear Metric Learning. In: Advances in Neural Information Processing Systems, vol 25, pp 2582–2590
Lacasse A, Laviolette F, Marchand M, Germain P, Usunier N (2007) PAC-Bayes Bounds for the Risk of the Majority Vote and the Variance of the Gibbs Classifier. In: Advances in Neural Information Processing Systems
Laviolette F, Marchand M (2007) PAC-Bayes Risk Bounds for Stochastic Averages and Majority Votes of Sample-Compressed Classifiers. Journal of Machine Learning Research 8:1461–1487
Laviolette F, Marchand M, Roy JF (2011) From PAC-Bayes Bounds to Quadratic Programs for Majority Votes. In: International Conference on Machine Learning
Lever G, Laviolette F, Shawe-Taylor J (2013) Tighter pac-bayes bounds through distribution-dependent priors. Theoretical Computer Science 473:4–28
Marchand M, Sokolova M (2005) Learning with Decision Lists of Data-Dependent Features. Journal of Machine Learning Research 6:427–451
Maurer A (2004) A Note on the PAC Bayesian Theorem. CoRR cs.LG/0411099
McAllester DA (1999) PAC-Bayesian model averaging. In: Annual Conference on Learning Theory, pp 164–170
McAllester DA (2003) Simplified PAC-Bayesian margin bounds. In: Annual Conference on Learning Theory, pp 203–215
Morvant E, Koço S, Ralaivola L (2012) PAC-Bayesian Generalization Bound on Confusion Matrix for Multi-Class Classification. In: International Conference on Machine Learning
Morvant E, Habrard A, Ayache S (2014) Majority Vote of Diverse Classifiers for Late Fusion. In: IAPR Joint International Workshops on Statistical Techniques in Pattern Recognition and Structural and Syntactic Pattern Recognition
Nock R, Sebban M, Bernard D (2003) A Simple Locally Adaptive Nearest Neighbor Rule With Application To Pollution Forecasting. International Journal of Pattern Recognition and Artificial Intelligence 17(8):1369–1382
Opelt A, Fussenegger M, Pinz A, Auer P (2004) Weak Hypotheses and Boosting for Generic Object Detection and Recognition. In: European Conference on Computer Vision, pp 71–84
Re M, Valentini G (2012) Ensemble methods: a review. Advances in machine learning and data mining for astronomy pp 563–582
Schapire R, Singer Y (1999) Improved boosting using confidence-rated predictions. Machine Learning 37(3):297336
Sun S (2013) A survey of multi-view machine learning. Neural Computing and Applications 23(7-8):2031–2038
Weinberger KQ, Saul LK (2009) Distance Metric Learning for Large Margin Nearest Neighbor Classification. Journal of Machine Learning Research 10:207–244
Xu C, Tao D, Xu C (2013) A Survey on Multi-view Learning. Tech. rep., arXiv:1304.5634

A Appendices

A.1 Tools

Theorem 4 (Markov's inequality) *Let Z be a random variable and $t \geq 0$, then: $P(|Z| \geq t) \leq \mathbf{E}(|Z|)/t$.*

Theorem 5 (Jensen's inequality) *Let X be an integrable real-valued random variable and $g(\cdot)$ convex, then: $g(\mathbf{E}[Z]) \leq \mathbf{E}[g(Z)]$.*

Lemma 1 (from inequalities (1) and (2) of Maurer (2004)) *Let $m \geq 8$, and $X = (X_1, .., X_m)$ be a vector of i.i.d. random variables, $0 \leq X_i \leq 1$. Then: $\sqrt{m} \leq \mathbf{E} \exp(m \, \mathrm{kl}(\frac{1}{m}\sum_{i=1}^{n} X_i \| \mathbf{E}[X_i])) \leq 2\sqrt{m}$, where $\mathrm{kl}(a\|b) = a \ln \frac{a}{b} + (1-a) \ln \frac{1-a}{1-b}$.*

A.2 Proof of Proposition 1

We give here another version of the proof of Proposition 4 of Germain et al. (2011).
Let Q be a distribution over $\mathcal{H}$, let $M = \max_{k' \in \{1,\ldots,n\}} \frac{1}{P_{k'}} |Q_{k'+n} - Q_{k'}|$, and let Q' be defined as $Q'_k = \frac{P_k}{2} + \frac{Q_k - Q_{k+n}}{2M}$, where by convention $(k+n)+n = k$ and $P_{k+n} = P_k$. First, let us show that Q' is actually $\mathbf{P}$-aligned on the auto-complemented $\mathcal{H}$, that is $\forall k \in \{1, \ldots, n\}$, $Q'_k \leq P_k$ and $Q'_k + Q'_{k+n} = P_k$. The following always holds:

$$
\begin{aligned}
Q'_k \leq P_k \iff \frac{P_k}{2} + \frac{Q_k - Q_{k+n}}{2M} \leq P_k \iff \frac{Q_k - Q_{k+n}}{M} &\leq P_k \\
\iff \frac{1}{P_k}(Q_k - Q_{k+n}) &\leq \max_{k' \in \{1,\ldots,n\}} \frac{1}{P_{k'}} |Q_{k'+n} - Q_{k'}|, \\
\text{and :} \quad Q'_k + Q'_{k+n} &= \frac{P_k}{2} + \frac{Q_k - Q_{k+n}}{2M} + \frac{P_{k+n}}{2} + \frac{Q_{k+n} - Q_k}{2M} \\
&= P_k + \frac{Q_k - Q_{k+n} + Q_{k+n} - Q_k}{2M} \;=\; P_k.
\end{aligned}
$$

Then, let us show that using Q' does not restrict the set of possible majority votes:

$$
\begin{aligned}
\mathop{\mathbf{E}}_{h \sim Q'} h(x) &= \sum_{k=1}^{2n} Q'_k h_k(\mathbf{x}) = \sum_{k=1}^{n} (Q'_k - Q'_{k+n}) h_k(\mathbf{x}) \\
&= \frac{1}{M} \sum_{k=1}^{n} (Q_k - Q_{k+n}) h_k(\mathbf{x}) = \frac{1}{M} \sum_{k=1}^{2n} Q_k h_k(\mathbf{x}) = \frac{1}{M} \mathop{\mathbf{E}}_{h \sim Q} h(\mathbf{x}).
\end{aligned}
$$

Therefore, we deduce that $\forall \mathbf{x} \in \mathcal{X}$, $B_{Q'}(\mathbf{x}) = B_Q(\mathbf{x})$ and since the constant term $\frac{1}{M}$ is present in both first and second moments $\mathcal{M}^D_{Q'}$ and $\mathcal{M}^D_{Q'^2}$, it vanishes in the C-bound. Hence, $C^D_{Q'} = C^D_Q$ regardless of the distribution D over $\mathcal{X} \times \mathcal{Y}$.

A.3 Proof of Algorithm 2 : P-MinCq

The Objective Function. We show how to obtain Line (6) from the definition of $\mathcal{M}^S_{Q^2}$.

$$
\begin{aligned}
\mathcal{M}^S_{Q^2} &= \mathop{\mathbf{E}}_{(h,h') \sim Q^2} \mathcal{M}^S_{h,h'} = \sum_{k=1}^{2n} \sum_{k'=1}^{2n} Q_k Q_{k'} \mathcal{M}^S_{h_k, h_{k'}} \\
&= \sum_{k=1}^{n} \sum_{k'=1}^{n} \Big[Q_k Q_{k'} \mathop{\mathbf{E}}_{(\mathbf{x},y) \sim S} h_k(\mathbf{x}) h_{k'}(\mathbf{x}) + Q_{k+n} Q_{k'} \mathop{\mathbf{E}}_{(\mathbf{x},y) \sim S} h_{k+n}(\mathbf{x}) h_{k'}(\mathbf{x}) \\
&\qquad + Q_k Q_{k'+n} \mathop{\mathbf{E}}_{(\mathbf{x},y) \sim S} h_k(\mathbf{x}) h_{k'+n}(\mathbf{x}) + Q_{k+n} Q_{k'+n} \mathop{\mathbf{E}}_{(\mathbf{x},y) \sim S} h_{k+n}(\mathbf{x}) h_{k'+n}(\mathbf{x}) \Big] \\
&= \sum_{k=1}^{n} \sum_{k'=1}^{n} Q_k Q_{k'} \mathop{\mathbf{E}}_{(\mathbf{x},y) \sim S} h_k(\mathbf{x}) h_{k'}(\mathbf{x}) - Q_{k+n} Q_{k'} \mathop{\mathbf{E}}_{(\mathbf{x},y) \sim S} h_k(\mathbf{x}) h_{k'}(\mathbf{x}) \\
&\qquad - Q_k Q_{k'+n} \mathop{\mathbf{E}}_{(\mathbf{x},y) \sim S} h_k(\mathbf{x}) h_{k'}(\mathbf{x}) + Q_{k+n} Q_{k'+n} \mathop{\mathbf{E}}_{(\mathbf{x},y) \sim S} h_k(\mathbf{x}) h_{k'}(\mathbf{x}) \;\; \text{(because } h_{k+n} = -h_k\text{)} \\
&= \sum_{k=1}^{n} \sum_{k'=1}^{n} \mathcal{M}^S_{h_k, h_{k'}} [Q_k Q_{k'} - (P_k - Q_k) Q_{k'} - Q_k (P_{k'} - Q_{k'}) + (P_k - Q_k)(P_{k'} - Q_{k'})]
\end{aligned}
$$

$$= \sum_{k=1}^{n}\sum_{k'=1}^{n} \mathcal{M}^S_{h_k,h_{k'}}[4Q_kQ_{k'} - 2P_kQ_{k'} - 2P_{k'}Q_k + P_kP_{k'}]$$
$$= 4\sum_{k=1}^{n}\sum_{k'=1}^{n} Q_k\mathcal{M}^S_{h_k,h_{k'}}Q_{k'} - 4\sum_{k=1}^{n}\sum_{k'=1}^{n} P_k\mathcal{M}^S_{h_k,h_{k'}}Q_{k'} + \sum_{k=1}^{n}\sum_{k'=1}^{n} P_kP_{k'}\mathcal{M}^S_{h_k,h_{k'}}$$
$$= 4[(\mathbf{Q}-\mathbf{P})^{\mathbf{T}}\mathbf{M_S}\mathbf{Q}] + C_1,$$

where $C_1 = \sum_{k=1}^{n}\sum_{k'=1}^{n} P_kP_{k'}\mathcal{M}_{h_k,h_{k'}}$ and the multiplicative value 4 can be considered as constant *w.r.t.* Q. Therefore, we get Line (6) of the optimization problem.

The Margin Constraint. We now show how to obtain Line (7) from $\mathcal{M}^{\mathcal{S}}_{Q}$. We have:

$$\mathcal{M}^{\mathcal{S}}_{Q} = \mathop{\mathbf{E}}_{h\sim Q}\mathcal{M}^S_h = \sum_{k=1}^{2n} Q_k\mathcal{M}^S_{h_k} = \sum_{k=1}^{n}(Q_k - Q_{k+n})\mathcal{M}^S_{h_k} = \sum_{k=1}^{n}(2Q_k - P_k)\mathcal{M}^S_{h_k} = \mathbf{m}^{\mathbf{T}}_{\mathbf{S}}(2\mathbf{Q}-\mathbf{P}),$$

where $\mathbf{m}^{\mathbf{T}}_{\mathbf{S}} = (\mathcal{M}_{h_1},\ldots,\mathcal{M}_{h_n})^T$. Replacing $\mathcal{M}^{\mathcal{S}}_{Q}$ by μ, we get Line (7) of the optimization problem.

A.4 Proof of Theorem 3

Proof of Equation (9)**.** Let S be any training sequence of size m. Suppose that $\mathcal{H}^S$ is auto-complemented. Moreover, a distribution on $\mathcal{H}^S$ is **P**-aligned if for any $(\mathbf{j},\sigma) \in \mathbf{J}_m \times \Omega_{S_\mathbf{j}}$ we have:

$$Q\left(h^{(\sigma,+)}_S\right) + Q\left(-h^{(\sigma,+)}_S\right) = Q\left(h^{(\sigma,+)}_S\right) + Q\left(h^{(\sigma,-)}_S\right) = P\left(h^{(\sigma,+)}_S\right) + P\left(h^{(\sigma,-)}_S\right) = P\left(h^{(\sigma,+)}_S\right) + P\left(-h^{(\sigma,+)}_S\right).$$

It implies that: $\mathcal{M}^D_{h^{(\sigma,+)}_S} = -\mathcal{M}^D_{h^{(\sigma,-)}_S}$, and:

$$\left(\mathcal{M}^S_{h^{(\sigma,+)}_{S_\mathbf{j}}} - \mathcal{M}^D_{h^{(\sigma,+)}_{S_\mathbf{j}}}\right)^2 = \left(-\mathcal{M}^S_{h^{(\sigma,-)}_{S_\mathbf{j}}} - (-\mathcal{M}^D_{h^{(\sigma,-)}_{S_\mathbf{j}}})\right)^2 = \left(\mathcal{M}^S_{h^{(\sigma,-)}_{S_\mathbf{j}}} - \mathcal{M}^D_{h^{(\sigma,-)}_{S_\mathbf{j}}}\right)^2.$$

Similarly as in McAllester (2003), we now consider the following Laplace transform:

$$X_P = \mathop{\mathbf{E}}_{h^{\omega}_{S_\mathbf{j}}\sim P} \exp\left(\frac{m-|\mathbf{j}|}{2B^2}(\mathcal{M}^S_{h^{\omega}_{S_\mathbf{j}}} - \mathcal{M}^D_{h^{\omega}_{S_\mathbf{j}}})^2\right).$$

Remark that $f(a,b) = \frac{1}{2B^2}(a-b)^2$ is convex because its Hessian matrix is positive semi-definite. For lightening the proof reading, we denote $m_\mathbf{j} = \dfrac{m-|\mathbf{j}|}{2B^2}$. For any **P**-aligned distribution Q, we have:

$$\begin{aligned}
2X_P &= \mathop{\mathbf{E}}_{h^{\omega}_{S_\mathbf{j}}\sim P} \exp\left(m_\mathbf{j}(\mathcal{M}^S_{h^{\omega}_{S_\mathbf{j}}} - \mathcal{M}^D_{h^{\omega}_{S_\mathbf{j}}})^2\right)\\
&= \int_{h^{(\sigma,+)}_{S_\mathbf{j}}\in\mathcal{H}^S} P(h^{(\sigma,+)}_{S_\mathbf{j}}) \exp\left(m_\mathbf{j}(\mathcal{M}^S_{h^{(\sigma,+)}_{S_\mathbf{j}}} - \mathcal{M}^D_{h^{(\sigma,+)}_{S_\mathbf{j}}})^2\right) dh^{(\sigma,+)}_{S_\mathbf{j}}\\
&\quad + \int_{h^{(\sigma,-)}_{S_\mathbf{j}}\in\mathcal{H}^S} P(h^{(\sigma,-)}_{S_\mathbf{j}}) \exp\left(m_\mathbf{j}(\mathcal{M}^S_{h^{(\sigma,-)}_{S_\mathbf{j}}} - \mathcal{M}^D_{h^{(\sigma,-)}_{S_\mathbf{j}}})^2\right) dh^{(\sigma,-)}_{S_\mathbf{j}}\\
&= \int_{h^{(\sigma,+)}_{S_\mathbf{j}}\in\mathcal{H}^S} \big(P(h^{(\sigma,+)}_{S_\mathbf{j}}) + P(-h^{(\sigma,+)}_{S_\mathbf{j}})\big) \exp\left(m_\mathbf{j}(\mathcal{M}^S_{h^{(\sigma,+)}_{S_\mathbf{j}}} - \mathcal{M}^D_{h^{(\sigma,+)}_{S_\mathbf{j}}})^2\right) dh^{(\sigma,+)}_{S_\mathbf{j}}\\
&= \int_{h^{(\sigma,+)}_{S_\mathbf{j}}\in\mathcal{H}^S} \big(Q(h^{(\sigma,+)}_{S_\mathbf{j}}) + Q(-h^{(\sigma,+)}_{S_\mathbf{j}})\big) \exp\left(m_\mathbf{j}(\mathcal{M}^S_{h^{(\sigma,+)}_{S_\mathbf{j}}} - \mathcal{M}^D_{h^{(\sigma,+)}_{S_\mathbf{j}}})^2\right) dh^{(\sigma,+)}_{S_\mathbf{j}}\\
&= \int_{h^{(\sigma,+)}_{S_\mathbf{j}}\in\mathcal{H}^S} Q(h^{(\sigma,+)}_{S_\mathbf{j}}) \exp\left(m_\mathbf{j}(\mathcal{M}^S_{h^{(\sigma,+)}_{S_\mathbf{j}}} - \mathcal{M}^D_{h^{(\sigma,+)}_{S_\mathbf{j}}})^2\right) dh^{(\sigma,+)}_{S_\mathbf{j}}\\
&\quad + \int_{h^{(\sigma,-)}_{S_\mathbf{j}}\in\mathcal{H}^S} Q(h^{(\sigma,-)}_{S_\mathbf{j}}) \exp\left(m_\mathbf{j}(\mathcal{M}^S_{h^{(\sigma,-)}_{S_\mathbf{j}}} - \mathcal{M}^D_{h^{(\sigma,-)}_{S_\mathbf{j}}})^2\right) dh^{(\sigma,-)}_{S_\mathbf{j}}\\
&= 2\mathop{\mathbf{E}}_{h^{\omega}_{S_\mathbf{j}}\sim Q} \exp\left(m_\mathbf{j}(\mathcal{M}^S_{h^{\omega}_{S_\mathbf{j}}} - \mathcal{M}^D_{h^{\omega}_{S_\mathbf{j}}})^2\right) = 2X_Q.
\end{aligned}$$

Using Markov's inequality (Theorem 4) we have: $\mathop{\mathbf{Pr}}_{S\sim D^m}\left(X_P \leq \frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m} X_P\right) \geq 1-\delta$.

Taking the logarithm on each side of the innermost inequality, for any $\delta \in (0,1]$, with a probability at least $1-\delta$ over the choice of $S\sim D^m$, for all $\mathbf{P}$-aligned distribution Q on $\mathcal{H}^S$, we get:

$$\ln\left[\mathop{\mathbf{E}}_{h^\omega_{S_\mathbf{j}}\sim Q} \exp\left(m_\mathbf{j}(\mathcal{M}^S_{h^\omega_{S_\mathbf{j}}} - \mathcal{M}^D_{h^\omega_{S_\mathbf{j}}})^2\right)\right] \leq \ln\left[\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m} X_P\right].$$

We apply Jensen's inequality (Theorem 5) on the concave function $\ln(\cdot)$:

$$\ln\left[\mathop{\mathbf{E}}_{h^\omega_{S_\mathbf{j}}\sim Q} \exp\left(m_\mathbf{j}(\mathcal{M}^S_{h^\omega_{S_\mathbf{j}}} - \mathcal{M}^D_{h^\omega_{S_\mathbf{j}}})^2\right)\right] \geq \mathop{\mathbf{E}}_{h^\omega_{S_\mathbf{j}}\sim Q} m_\mathbf{j}(\mathcal{M}^S_{h^\omega_{S_\mathbf{j}}} - \mathcal{M}^D_{h^\omega_{S_\mathbf{j}}})^2.$$

Recall that $|\mathbf{j}^{\max}|$ is the maximal size of the compression sample. Then by again applying the Jensen's inequality on the convex function $(m-|\mathbf{j}^{\max}|)f(a,b) = \frac{m-|\mathbf{j}^{\max}|}{2B^2}(a-b)^2 = m_\mathbf{j}(a-b)^2$ for the left side of the previous inequality, we have:

$$\begin{aligned}
\mathop{\mathbf{E}}_{h^\omega_{S_\mathbf{j}}\sim Q} m_\mathbf{j}(\mathcal{M}^S_{h^\omega_{S_\mathbf{j}}} - \mathcal{M}^D_{h^\omega_{S_\mathbf{j}}})^2 &= \frac{m}{2B^2}\left(\mathop{\mathbf{E}}_{h^\omega_{S_\mathbf{j}}\sim Q} - |\mathbf{j}|(\mathcal{M}^S_{h^\omega_{S_\mathbf{j}}} - \mathcal{M}^D_{h^\omega_{S_\mathbf{j}}})^2\right)\\
&\geq \frac{m-|\mathbf{j}^{\max}|}{2B^2}\left(\mathop{\mathbf{E}}_{h^\omega_{S_\mathbf{j}}\sim Q}(\mathcal{M}^S_{h^\omega_{S_\mathbf{j}}} - \mathcal{M}^D_{h^\omega_{S_\mathbf{j}}})^2\right)\\
&\geq \frac{m-|\mathbf{j}^{\max}|}{2B^2}(\mathcal{M}^S_Q - \mathcal{M}^D_Q)^2.
\end{aligned}$$

Then: $\mathop{\mathbf{Pr}}_{S\sim D^m}\left(\frac{m-|\mathbf{j}^{\max}|}{2B^2}\left(\mathcal{M}^S_Q - \mathcal{M}^D_Q\right)^2 \leq \ln\left[\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m} X_P\right]\right) \geq 1-\delta$.

We thus have to bound $\mathop{\mathbf{E}}_{S\sim D^m} X_P$. We consider $\mathcal{M}^{S\setminus S_\mathbf{j}}_{h^\omega_{S_\mathbf{j}}}$ the empirical margin computed on the examples of the learning sample S that are not in the compression sequence $S_\mathbf{j}$. While $\mathcal{M}^S_{h^\omega_{S_\mathbf{j}}}$ may contain some bias, $\mathcal{M}^{S\setminus S_\mathbf{j}}_{h^\omega_{S_\mathbf{j}}}$ is an arithmetic mean of truly *i.i.d.* $(m-|\mathbf{j}|)$ random variables. Note also that these two random variables have very close values.

We have: $0 \leq m\mathcal{M}^S_{h^\omega_{S_\mathbf{j}}} - (m-|\mathbf{j}|)\mathcal{M}^{S\setminus S_\mathbf{j}}_{h^\omega_{S_\mathbf{j}}} \leq B|\mathbf{j}|$,

then : $-B|\mathbf{j}| \leq -|\mathbf{j}|\mathcal{M}^{S\setminus S_\mathbf{j}}_{h^\omega_{S_\mathbf{j}}} \leq m\mathcal{M}^S_{h^\omega_{S_\mathbf{j}}} - m\mathcal{M}^{S\setminus S_\mathbf{j}}_{h^\omega_{S_\mathbf{j}}} \leq |\mathbf{j}| - |\mathbf{j}|\mathcal{M}^{S\setminus S_\mathbf{j}}_{h^\omega_{S_\mathbf{j}}} \leq B|\mathbf{j}|$,

and thus : $\left|\mathcal{M}^S_{h^\omega_{S_\mathbf{j}}} - \mathcal{M}^{S\setminus S_\mathbf{j}}_{h^\omega_{S_\mathbf{j}}}\right| \leq \frac{B|\mathbf{j}|}{m}$.

Given a compression sequence $S_\mathbf{j}$, we denote by $\bar{\mathbf{j}}$ the vector of indices that are not in $\mathbf{j}$. Then:

$$\begin{aligned}
\mathop{\mathbf{E}}_{S\sim D^m} X_P &= \mathop{\mathbf{E}}_{S\sim D^m}\mathop{\mathbf{E}}_{h^\omega_{S_\mathbf{j}}\sim P} \exp\left(m_\mathbf{j}(\mathcal{M}^S_{h^\omega_{S_\mathbf{j}}} - \mathcal{M}^D_{h^\omega_{S_\mathbf{j}}})^2\right)\\
&= \mathop{\mathbf{E}}_{\mathbf{j}\sim P}\mathop{\mathbf{E}}_{S_\mathbf{j}\sim D^{|\mathbf{j}|}}\mathop{\mathbf{E}}_{\omega\sim P_{S_\mathbf{j}}}\mathop{\mathbf{E}}_{S_{\bar{\mathbf{j}}}\sim D^{m-|\mathbf{j}|}} \exp\left(m_\mathbf{j}(\mathcal{M}^S_{h^\omega_{S_\mathbf{j}}} - \mathcal{M}^D_{h^\omega_{S_\mathbf{j}}})^2\right).
\end{aligned}$$

For all $\mathbf{j}\in\mathbf{J}_m$, $S_\mathbf{j}\in\mathcal{Z}^{|\mathbf{j}|}$, $\omega\in\Omega'_{S_\mathbf{j}}\times\{+,-\}$, we have :

$$\begin{aligned}
&\mathop{\mathbf{E}}_{S\sim D^m} X_P = \mathop{\mathbf{E}}_{S_{\bar{\mathbf{j}}}\sim D^{m-|\mathbf{j}|}} \exp\left(m_\mathbf{j}(\mathcal{M}^S_{h^\omega_{S_\mathbf{j}}} - \mathcal{M}^D_{h^\omega_{S_\mathbf{j}}})^2\right)\\
&= \mathop{\mathbf{E}}_{S_{\bar{\mathbf{j}}}\sim D^{m-|\mathbf{j}|}} \exp\left(m_\mathbf{j}(\mathcal{M}^S_{h^\omega_{S_\mathbf{j}}} - \mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^\omega_{S_\mathbf{j}}} + \mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^\omega_{S_\mathbf{j}}} - \mathcal{M}^D_{h^\omega_{S_\mathbf{j}}})^2\right)\\
&\leq \mathop{\mathbf{E}}_{S_{\bar{\mathbf{j}}}\sim D^{m-|\mathbf{j}|}} \exp\left[m_\mathbf{j}\left(\left[\mathcal{M}^S_{h^\omega_{S_\mathbf{j}}} - \mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^\omega_{S_\mathbf{j}}}\right]^2 + 2\left|\mathcal{M}^S_{h^\omega_{S_\mathbf{j}}} - \mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^\omega_{S_\mathbf{j}}}\right|\left|\mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^\omega_{S_\mathbf{j}}} - \mathcal{M}^D_{h^\omega_{S_\mathbf{j}}}\right| + \left[\mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^\omega_{S_\mathbf{j}}} - \mathcal{M}^D_{h^\omega_{S_\mathbf{j}}}\right]^2\right)\right].
\end{aligned}$$

From Equation (A.4) and since $\exp(.)$ is increasing, we obtain:

$$\mathop{\mathbf{E}}_{S\sim D^m} X_P \leq \mathop{\mathbf{E}}_{S_{\bar{\mathbf{j}}}\sim D^{m-|\mathbf{j}|}} \exp\left[m_{\mathbf{j}}\left(\left[\frac{B|\mathbf{j}|}{m}\right]^2 + 2\frac{B|\mathbf{j}|}{m} + [\mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^\omega_{S_{\mathbf{j}}}} - \mathcal{M}^{D}_{h^\omega_{S_{\mathbf{j}}}}]^2\right)\right].$$

Since we suppose that for all $\mathbf{j}$: $|\mathbf{j}| \leq |\mathbf{j}^{\max}| \leq m$, then:

$$\frac{m-|\mathbf{j}|}{2B}\left(\left[\frac{|\mathbf{j}|}{m}\right]^2 + 2\frac{|\mathbf{j}|}{m}\right) \leq |\mathbf{j}^{\max}|\left(\frac{m-|\mathbf{j}|}{2B}\left[\frac{|\mathbf{j}|}{m^2} + \frac{2}{m}\right]\right) \leq \frac{|\mathbf{j}^{\max}|}{B}.$$

Then:
$$\begin{aligned}
\mathop{\mathbf{E}}_{S\sim D^m} X_P &\leq \mathop{\mathbf{E}}_{S_{\bar{\mathbf{j}}}\sim D^{m-|\mathbf{j}|}} \exp\left(\frac{|\mathbf{j}^{\max}|}{B} + m_{\mathbf{j}}(\mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^\omega_{S_{\mathbf{j}}}} - \mathcal{M}^{D}_{h^\omega_{S_{\mathbf{j}}}})^2\right)\\
&\leq \exp\left(\frac{|\mathbf{j}^{\max}|}{B}\right) \times \mathop{\mathbf{E}}_{S_{\bar{\mathbf{j}}}\sim D^{m-|\mathbf{j}|}} \exp\left(m_{\mathbf{j}}[\mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^\omega_{S_{\mathbf{j}}}} - \mathcal{M}^{D}_{h^\omega_{S_{\mathbf{j}}}}]^2\right)\\
&\leq \exp\left(\frac{|\mathbf{j}^{\max}|}{B}\right) \times \mathop{\mathbf{E}}_{S_{\bar{\mathbf{j}}}\sim D^{m-|\mathbf{j}|}} \exp\left(2(m-|\mathbf{j}|)\left[\left(\frac{1}{2} - \frac{\mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^\omega_{S_{\mathbf{j}}}}}{2B}\right) - \left(\frac{1}{2} - \frac{\mathcal{M}^{D}_{h^\omega_{S_{\mathbf{j}}}}}{2B}\right)\right]^2\right).
\end{aligned}$$

By definition $2(a-b)^2 \leq \mathrm{kl}(a\|b) = a\ln\frac{a}{b} + (1-a)\ln\frac{1-a}{1-b}$ is valid for any $a, b \in [0,1]$ provided that if $a = 0$ then so is b and if $a = 1$ then so is b. Since the elements of $\mathcal{H}^S$ are B-bounded and $S_{\bar{\mathbf{j}}}$ is drawn *i.i.d.* from D, we have: $\mathcal{M}^{D}_{h^\omega_{S_{\mathbf{j}}}} = -B \Rightarrow \mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^\omega_{S_{\mathbf{j}}}} = -B$, and $\mathcal{M}^{D}_{h^\omega_{S_{\mathbf{j}}}} = B \Rightarrow \mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^\omega_{S_{\mathbf{j}}}} = B$.

Then: $\frac{1}{2} - \frac{\mathcal{M}^{D}_{h^\omega_{S_{\mathbf{j}}}}}{2B} = 0 \Rightarrow \frac{1}{2} - \frac{\mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^\omega_{S_{\mathbf{j}}}}}{2B} = 0$, and $\frac{1}{2} - \frac{\mathcal{M}^{D}_{h^\omega_{S_{\mathbf{j}}}}}{2B} = 1 \Rightarrow \frac{1}{2} - \frac{\mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^\omega_{S_{\mathbf{j}}}}}{2B} = 1$.

Moreover since: $0 \leq \frac{1}{2} - \frac{\mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^\omega_{S_{\mathbf{j}}}}}{2B} \leq 1$, and $0 \leq \frac{1}{2} - \frac{\mathcal{M}^{D}_{h^\omega_{S_{\mathbf{j}}}}}{2B} \leq 1$, we have:

$$\mathop{\mathbf{E}}_{S\sim D^m} X_P \leq \exp\left(\frac{|\mathbf{j}^{\max}|}{B}\right) \times \mathop{\mathbf{E}}_{S_{\bar{\mathbf{j}}}\sim D^{m-|\mathbf{j}|}} \exp\left((m-|\mathbf{j}|)\,\mathrm{kl}\left(\frac{1}{2} - \frac{\mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^\omega_{S_{\mathbf{j}}}}}{2B}\,\middle\|\,\frac{1}{2} - \frac{\mathcal{M}^{D}_{h^\omega_{S_{\mathbf{j}}}}}{2B}\right)\right).$$

We apply Maurer's Lemma (Lemma 1):

$$\begin{aligned}
\mathop{\mathbf{E}}_{S\sim D^m} X_P &\leq \exp\left(\frac{|\mathbf{j}^{\max}|}{B}\right) \times \mathop{\mathbf{E}}_{S_{\bar{\mathbf{j}}}\sim D^{m-|\mathbf{j}|}} 2\sqrt{(m-|\mathbf{j}|)}\\
&\leq \exp\left(\frac{|\mathbf{j}^{\max}|}{B}\right) \times 2\sqrt{(m-|\mathbf{j}|)} \leq \exp\left(\frac{|\mathbf{j}^{\max}|}{B}\right) \times 2\sqrt{m}.
\end{aligned}$$

Finally:
$$\mathop{\mathbf{Pr}}_{S\sim D^m}\left(\begin{array}{l}\text{for all } \mathbf{P}\text{-aligned distribution } Q \text{ on } \mathcal{H}^S,\\ |\mathcal{M}^D_Q - \mathcal{M}^S_Q| \leq \dfrac{2B\sqrt{\dfrac{|\mathbf{j}^{\max}|}{B} + \ln\left(\dfrac{2\sqrt{m}}{\delta}\right)}}{\sqrt{2(m-|\mathbf{j}^{\max}|)}}\end{array}\right) \geq 1-\delta$$

Proof of Equation (10)**.** Using similar arguments as the beginning of the proof Equation (9), we have:

$$\begin{aligned}
(\mathcal{M}^{S}_{h^{(\sigma,+)}_{S_{\mathbf{j}}}, h^{(\sigma,+)}_{S_{\mathbf{j}'}}} - \mathcal{M}^{D}_{h^{(\sigma,+)}_{S_{\mathbf{j}}}, h^{(\sigma,+)}_{S_{\mathbf{j}'}}})^2 &= (\mathcal{M}^{S}_{h^{(\sigma,-)}_{S_{\mathbf{j}}}, h^{(\sigma,+)}_{S_{\mathbf{j}'}}} - \mathcal{M}^{D}_{h^{(\sigma,-)}_{S_{\mathbf{j}}}, h^{(\sigma,+)}_{S_{\mathbf{j}'}}})^2\\
&= (\mathcal{M}^{S}_{h^{(\sigma,+)}_{S_{\mathbf{j}}}, h^{(\sigma,-)}_{S_{\mathbf{j}'}}} - \mathcal{M}^{D}_{h^{(\sigma,+)}_{S_{\mathbf{j}}}, h^{(\sigma,-)}_{S_{\mathbf{j}'}}})^2\\
&= (\mathcal{M}^{S}_{h^{(\sigma,-)}_{S_{\mathbf{j}'}}, h^{(\sigma,-)}_{S_{\mathbf{j}}}} - \mathcal{M}^{D}_{h^{(\sigma,-)}_{S_{\mathbf{j}'}}, h^{(\sigma,-)}_{S_{\mathbf{j}}}})^2.
\end{aligned}$$

Similarly as in McAllester (2003), we now consider the following Laplace transform:

$$X_P = \mathop{\mathbf{E}}_{h^{\omega}_{S_{\mathbf{j}}}, h^{\omega}_{S'_{\mathbf{j}}} \sim P^2} \exp\left(\frac{m - |\mathbf{j} \cup \mathbf{j}'|}{2B^4} \big(\mathcal{M}^S_{h^{\omega}_{S_{\mathbf{j}}}, h^{\omega}_{S_{\mathbf{j}'}}} - \mathcal{M}^D_{h^{\omega}_{S_{\mathbf{j}}}, h^{\omega'}_{S_{\mathbf{j}'}}}\big)^2 \right).$$

For lightening the proof reading, we denote $m_{\mathbf{j}\cup\mathbf{j}'} = \dfrac{m - |\mathbf{j} \cup \mathbf{j}'|}{2B^4}$. Remark that $f(a,b) = \frac{1}{2B^4}(a-b)^2$ is convex. For any $\mathbf{P}$-aligned distribution Q, we have:

$$\begin{aligned}
4X_P &= \mathop{\mathbf{E}}_{h^{\omega}_{S_{\mathbf{j}}}, h^{\omega}_{S'_{\mathbf{j}}} \sim P^2} \exp\Big(m_{\mathbf{j}\cup\mathbf{j}'}\big(\mathcal{M}^S_{h^{\omega}_{S_{\mathbf{j}}}, h^{\omega}_{S_{\mathbf{j}'}}} - \mathcal{M}^D_{h^{\omega}_{S_{\mathbf{j}}}, h^{\omega'}_{S_{\mathbf{j}'}}}\big)^2\Big) \\
&= \int_{h^{(\sigma,+)}_{S_{\mathbf{j}}}, h^{(\sigma,+)}_{S'_{\mathbf{j}}} \in (\mathcal{H}^S)^2} P(h^{(\sigma,+)}_{S_{\mathbf{j}}}) P(h^{(\sigma,+)}_{S_{\mathbf{j}'}}) \exp\Big(m_{\mathbf{j}\cup\mathbf{j}'}\big(\mathcal{M}^S_{h^{(\sigma,+)}_{S_{\mathbf{j}}}, h^{(\sigma,+)}_{S_{\mathbf{j}'}}} - \mathcal{M}^D_{h^{(\sigma,+)}_{S_{\mathbf{j}}}, h^{(\sigma,+)}_{S_{\mathbf{j}'}}}\big)^2\Big) dh^{(\sigma,+)}_{S_{\mathbf{j}}} h^{(\sigma,+)}_{S_{\mathbf{j}'}} \\
&+ \int_{h^{(\sigma,-)}_{S_{\mathbf{j}}}, h^{(\sigma,-)}_{S'_{\mathbf{j}}} \in (\mathcal{H}^S)^2} P(h^{(\sigma,-)}_{S_{\mathbf{j}}}) P(h^{(\sigma,-)}_{S_{\mathbf{j}'}}) \exp\Big(m_{\mathbf{j}\cup\mathbf{j}'}\big(\mathcal{M}^S_{h^{(\sigma,-)}_{S_{\mathbf{j}}}, h^{(\sigma,-)}_{S_{\mathbf{j}'}}} - \mathcal{M}^D_{h^{(\sigma,-)}_{S_{\mathbf{j}}}, h^{(\sigma,-)}_{S_{\mathbf{j}'}}}\big)^2\Big) dh^{(\sigma,-)}_{S_{\mathbf{j}}} h^{(\sigma,-)}_{S_{\mathbf{j}'}} \\
&+ \int_{h^{(\sigma,-)}_{S_{\mathbf{j}}}, h^{(\sigma,+)}_{S'_{\mathbf{j}}} \in (\mathcal{H}^S)^2} P(h^{(\sigma,-)}_{S_{\mathbf{j}}}) P(h^{(\sigma,+)}_{S_{\mathbf{j}'}}) \exp\Big(m_{\mathbf{j}\cup\mathbf{j}'}\big(\mathcal{M}^S_{h^{(\sigma,-)}_{S_{\mathbf{j}}}, h^{(\sigma,+)}_{S_{\mathbf{j}'}}} - \mathcal{M}^D_{h^{(\sigma,-)}_{S_{\mathbf{j}}}, h^{(\sigma,+)}_{S_{\mathbf{j}'}}}\big)^2\Big) dh^{(\sigma,-)}_{S_{\mathbf{j}}} h^{(\sigma,+)}_{S_{\mathbf{j}'}} \\
&+ \int_{h^{(\sigma,+)}_{S_{\mathbf{j}}}, h^{(\sigma,-)}_{S'_{\mathbf{j}}} \in (\mathcal{H}^S)^2} P(h^{(\sigma,+)}_{S_{\mathbf{j}}}) P(h^{(\sigma,-)}_{S_{\mathbf{j}'}}) \exp\Big(m_{\mathbf{j}\cup\mathbf{j}'}\big(\mathcal{M}^S_{h^{(\sigma,+)}_{S_{\mathbf{j}}}, h^{(\sigma,-)}_{S_{\mathbf{j}'}}} - \mathcal{M}^D_{h^{(\sigma,+)}_{S_{\mathbf{j}}}, h^{(\sigma,-)}_{S_{\mathbf{j}'}}}\big)^2\Big) dh^{(\sigma,+)}_{S_{\mathbf{j}}} h^{(\sigma,-)}_{S_{\mathbf{j}'}} \\
&= \int_{h^{(\sigma,+)}_{S_{\mathbf{j}}}, h^{(\sigma,+)}_{S'_{\mathbf{j}}} \in (\mathcal{H}^S)^2} \big(P(h^{(\sigma,+)}_{S_{\mathbf{j}}}) + P(-h^{(\sigma,+)}_{S_{\mathbf{j}}})\big)\big(P(h^{(\sigma,+)}_{S_{\mathbf{j}'}}) + P(-h^{(\sigma,+)}_{S_{\mathbf{j}'}})\big) \\
&\qquad \exp\Big(m_{\mathbf{j}\cup\mathbf{j}'}\big(\mathcal{M}^S_{h^{(\sigma,+)}_{S_{\mathbf{j}}}, h^{(\sigma,+)}_{S_{\mathbf{j}'}}} - \mathcal{M}^D_{h^{(\sigma,+)}_{S_{\mathbf{j}}}, h^{(\sigma,+)}_{S_{\mathbf{j}'}}}\big)^2\Big) dh^{(\sigma,+)}_{S_{\mathbf{j}}} h^{(\sigma,+)}_{S_{\mathbf{j}'}} \\
&= \int_{h^{(\sigma,+)}_{S_{\mathbf{j}}}, h^{(\sigma,+)}_{S'_{\mathbf{j}}} \in (\mathcal{H}^S)^2} \big(Q(h^{(\sigma,+)}_{S_{\mathbf{j}}}) + Q(-h^{(\sigma,+)}_{S_{\mathbf{j}}})\big)\big(Q(h^{(\sigma,+)}_{S_{\mathbf{j}'}}) + Q(-h^{(\sigma,+)}_{S_{\mathbf{j}'}})\big) \\
&\qquad \exp\Big(m_{\mathbf{j}\cup\mathbf{j}'}\big(\mathcal{M}^S_{h^{(\sigma,+)}_{S_{\mathbf{j}}}, h^{(\sigma,+)}_{S_{\mathbf{j}'}}} - \mathcal{M}^D_{h^{(\sigma,+)}_{S_{\mathbf{j}}}, h^{(\sigma,+)}_{S_{\mathbf{j}'}}}\big)^2\Big) dh^{(\sigma,+)}_{S_{\mathbf{j}}} h^{(\sigma,+)}_{S_{\mathbf{j}'}} \\
&= \cdots\cdots\cdots \\
&= 4 \mathop{\mathbf{E}}_{h^{\omega}_{S_{\mathbf{j}}}, h^{\omega}_{S'_{\mathbf{j}}} \sim Q^2} \exp\Big(m_{\mathbf{j}\cup\mathbf{j}'}\big(\mathcal{M}^S_{h^{\omega}_{S_{\mathbf{j}}}, h^{\omega}_{S_{\mathbf{j}'}}} - \mathcal{M}^D_{h^{\omega}_{S_{\mathbf{j}}}, h^{\omega'}_{S_{\mathbf{j}'}}}\big)^2\Big) = 4X_Q.
\end{aligned}$$

Now, by Markov's inequality (Theorem 4) we have: $\mathop{\mathbf{Pr}}_{S\sim D^m}\left(X_P \leq \dfrac{1}{\delta} \mathop{\mathbf{E}}_{S\sim D^m} X_P\right) \geq 1 - \delta.$
By taking the logarithm on each side of the innermost inequality, for any $\delta \in (0,1]$, with a probability at least $1-\delta$ over the choice of $S \sim D^m$, for all $\mathbf{P}$-aligned distribution Q on $\mathcal{H}^S$ we have:

$$\ln\left[\mathop{\mathbf{E}}_{h^{\omega}_{S_{\mathbf{j}}}, h^{\omega'}_{S_{\mathbf{j}'}} \sim Q^2} \exp\Big(m_{\mathbf{j}\cup\mathbf{j}'}\big(\mathcal{M}^S_{h^{\omega}_{S_{\mathbf{j}}}, h^{\omega'}_{S_{\mathbf{j}'}}} - \mathcal{M}^D_{h^{\omega}_{S_{\mathbf{j}}}, h^{\omega'}_{S_{\mathbf{j}'}}}\big)^2\Big)\right] \leq \ln\left[\frac{1}{\delta} \mathop{\mathbf{E}}_{S\sim D^m} X_P\right].$$

We apply Jensen's inequality (Theorem 5) on $\ln(.)$:

$$\ln\left[\mathop{\mathbf{E}}_{h^{\omega}_{S_{\mathbf{j}}}, h^{\omega'}_{S_{\mathbf{j}'}} \sim Q^2} \exp\left(m_{\mathbf{j}\cup\mathbf{j}'}\big(\mathcal{M}^S_{h^{\omega}_{S_{\mathbf{j}}}, h^{\omega'}_{S_{\mathbf{j}'}}} - \mathcal{M}^D_{h^{\omega}_{S_{\mathbf{j}}}, h^{\omega'}_{S_{\mathbf{j}'}}}\big)^2\right)\right] \geq \mathop{\mathbf{E}}_{h^{\omega}_{S_{\mathbf{j}}}, h^{\omega'}_{S_{\mathbf{j}'}} \sim Q^2} m_{\mathbf{j}\cup\mathbf{j}'}\big(\mathcal{M}^S_{h^{\omega}_{S_{\mathbf{j}}}, h^{\omega'}_{S_{\mathbf{j}'}}} - \mathcal{M}^D_{h^{\omega}_{S_{\mathbf{j}}}, h^{\omega'}_{S_{\mathbf{j}'}}}\big)^2.$$

Recall that $|\mathbf{j}^{\max}| < \frac{m}{2}$ the maximal size of the compression sample. Then by again applying the Jensen's inequality on the convex function $(m - |\mathbf{j}^{\max}|)f(a,b) = \frac{m-|\mathbf{j}^{\max}|}{2B^4}(a-b)^2 = m_{\mathbf{j}\cup\mathbf{j}'}(a-b)^2$ for the left

side of the previous inequality, we have:

$$\mathop{\mathbf{E}}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}\sim Q^2} m_{\mathbf{j}\cup\mathbf{j}'}\big(\mathcal{M}^{S}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}-\mathcal{M}^{D}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}\big)^2=\frac{m}{2B^4}\left(\mathop{\mathbf{E}}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}\sim Q^2}(-|\mathbf{j}\cup\mathbf{j}'|)\big(\mathcal{M}^{S}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}-\mathcal{M}^{D}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}\big)^2\right)$$

$$\geq \frac{m\ -2|\mathbf{j}^{\max}|}{2B^4}\left(\mathop{\mathbf{E}}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}\sim Q^2}\big(\mathcal{M}^{S}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}-\mathcal{M}^{D}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}\big)^2\right)$$

$$\geq \frac{m-2|\mathbf{j}^{\max}|}{2B^4}\big(\mathcal{M}^{S}_{Q^2}-\mathcal{M}^{D}_{Q^2}\big)^2.$$

Then: $\mathop{\mathbf{Pr}}_{S\sim D^m}\left(\frac{m-2|\mathbf{j}^{\max}|}{2B^4}\big(\mathcal{M}^{S}_{Q^2}-\mathcal{M}^{D}_{Q^2}\big)^2\leq \ln\left[\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m}X_P\right]\right)\geq 1-\delta.$

We thus have to bound $\mathop{\mathbf{E}}_{S\sim D^m}X_P$. We consider $\mathcal{M}^{S\setminus(S_{\mathbf{j}}\cup S_{\mathbf{j}'})}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}$ the empirical second moment of the margin computed on the examples of the learning sample S that are not in the compression sequence $S_{\mathbf{j}}$. While $\mathcal{M}^{S}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}$ may contain some bias, $\mathcal{M}^{S\setminus(S_{\mathbf{j}}\cup S_{\mathbf{j}'})}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}$ is an arithmetic mean of truly *i.i.d.* $(m-|\mathbf{j}\cup\mathbf{j}'|)$ random variables. We can also note that these two random variables have very close values. We have:

$$0\leq m\mathcal{M}^{S}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}-(m-|\mathbf{j}\cup\mathbf{j}'|)\mathcal{M}^{S\setminus(S_{\mathbf{j}}\cup S_{\mathbf{j}'})}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}\leq B^2|\mathbf{j}\cup\mathbf{j}'|,$$

then:

$$-B^2|\mathbf{j}\cup\mathbf{j}'|\leq -|\mathbf{j}\cup\mathbf{j}'|\mathcal{M}^{S\setminus(S_{\mathbf{j}}\cup S_{\mathbf{j}'})}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}\leq m\mathcal{M}^{S}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}-m\mathcal{M}^{S\setminus(S_{\mathbf{j}}\cup S_{\mathbf{j}'})}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}\leq |\mathbf{j}\cup\mathbf{j}'|-|\mathbf{j}\cup\mathbf{j}'|\mathcal{M}^{S\setminus(S_{\mathbf{j}}\cup S_{\mathbf{j}'})}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}\leq B^2|\mathbf{j}\cup\mathbf{j}'|,$$

thus:
$$\left|\mathcal{M}^{S}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}-\mathcal{M}^{S\setminus(S_{\mathbf{j}}\cup S_{\mathbf{j}'})}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}\right|\leq\frac{B^2|\mathbf{j}\cup\mathbf{j}'|}{m}. \tag{13}$$

Given two compression sequences $S_{\mathbf{j}}$ and $S_{\mathbf{j}'}$, Let $\bar{\mathbf{j}}$ be the vector of indices that are not in $\mathbf{j}\cup\mathbf{j}'$. Then:

$$\mathop{\mathbf{E}}_{S\sim D^m}X_P=\mathop{\mathbf{E}}_{S\sim D^m}\mathop{\mathbf{E}}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}\sim P^2}\exp\left(m_{\mathbf{j}\cup\mathbf{j}'}\big(\mathcal{M}^{S}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}-\mathcal{M}^{D}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}\big)^2\right)$$

$$=\mathop{\mathbf{E}}_{\mathbf{j},\mathbf{j}'\sim P^2}\mathop{\mathbf{E}}_{S_{\mathbf{j}},S_{\mathbf{j}'}\sim D^{|\mathbf{j}|}\times D^{|\mathbf{j}'|}}\mathop{\mathbf{E}}_{\omega,\omega'\sim P_{S_{\mathbf{j}}}\times P_{S_{\mathbf{j}'}}}\mathop{\mathbf{E}}_{S_{\bar{\mathbf{j}}}\sim D^{m-|\mathbf{j}\cup\mathbf{j}'|}}\exp\left(m_{\mathbf{j}\cup\mathbf{j}'}\big(\mathcal{M}^{S}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}-\mathcal{M}^{D}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}\big)^2\right).$$

For all $\mathbf{j},\mathbf{j}'\in(\mathbf{J}_m)^2$, $S_{\mathbf{j}},S_{\mathbf{j}'}\in\mathcal{Z}^{|\mathbf{j}|}\times\mathcal{Z}^{|\mathbf{j}'|}$, $\omega,\omega'\in(\Omega'_{S_{\mathbf{j}}}\times\{+,-\})\times(\Omega'_{S_{\mathbf{j}'}}\times\{+,-\})$, we have:

$$\mathop{\mathbf{E}}_{S_{\bar{\mathbf{j}}}\sim D^{m-|\mathbf{j}\cup\mathbf{j}'|}}\exp\left(m_{\mathbf{j}\cup\mathbf{j}'}\big(\mathcal{M}^{S}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}-\mathcal{M}^{D}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}\big)^2\right)$$

$$=\mathop{\mathbf{E}}_{S_{\bar{\mathbf{j}}}\sim D^{m-|\mathbf{j}\cup\mathbf{j}'|}}\exp\left(m_{\mathbf{j}\cup\mathbf{j}'}\big(\mathcal{M}^{S}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}-\mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}+\mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}-\mathcal{M}^{D}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}\big)^2\right)$$

$$\leq\mathop{\mathbf{E}}_{S_{\bar{\mathbf{j}}}\sim D^{m-|\mathbf{j}\cup\mathbf{j}'|}}\exp\left[m_{\mathbf{j}\cup\mathbf{j}'}\Big(\big[\mathcal{M}^{S}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}-\mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}\big]^2\right.$$
$$\left.+2\big|\mathcal{M}^{S}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}-\mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}\big|\big|\mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}-\mathcal{M}^{D}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}\big|+\big[\mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}-\mathcal{M}^{D}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}\big]^2\Big)\right].$$

From Equation (13), since $\exp(.)$ is increasing we obtain:

$$\mathop{\mathbf{E}}_{S\sim D^m}X_P\leq\mathop{\mathbf{E}}_{S_{\bar{\mathbf{j}}}\sim D^{m-|\mathbf{j}\cup\mathbf{j}'|}}\exp\left[m_{\mathbf{j}\cup\mathbf{j}'}\left(\left[\frac{B^2|\mathbf{j}\cup\mathbf{j}'|}{m}\right]^2+2\frac{B^2|\mathbf{j}\cup\mathbf{j}'|}{m}+\big[\mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}-\mathcal{M}^{D}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}\big]^2\right)\right].$$

Since we suppose that for all $\mathbf{j}$ we have $|\mathbf{j}| \leq |\mathbf{j}^{\max}| \leq \frac{m}{2}$, we can easily compute:

$$m_{\mathbf{j}\cup\mathbf{j}'}\left(\left[\tfrac{|\mathbf{j}\cup\mathbf{j}'|}{m}\right]^2 + 2\tfrac{|\mathbf{j}\cup\mathbf{j}'|}{m}\right) \leq 2|\mathbf{j}^{\max}|\left[m_{\mathbf{j}\cup\mathbf{j}'}\left(\tfrac{|\mathbf{j}\cup\mathbf{j}'|}{m^2} + \tfrac{2}{m}\right)\right] \leq \tfrac{2|\mathbf{j}^{\max}|}{B^2}.$$

Then:

$$\begin{aligned}
\mathop{\mathbf{E}}_{S\sim D^m} X_P &\leq \mathop{\mathbf{E}}_{S_{\bar{\mathbf{j}}}\sim D^{m-|\mathbf{j}\cup\mathbf{j}'|}} \exp\left[\frac{2|\mathbf{j}^{\max}|}{B^2} + m_{\mathbf{j}\cup\mathbf{j}'}\left[\mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}} - \mathcal{M}^{D}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}\right]^2\right] \\
&\leq \exp\left[\frac{2|\mathbf{j}^{\max}|}{B^2}\right] \mathop{\mathbf{E}}_{S_{\bar{\mathbf{j}}}\sim D^{m-|\mathbf{j}\cup\mathbf{j}'|}} \exp\left[m_{\mathbf{j}\cup\mathbf{j}'}\left[\mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}} - \mathcal{M}^{D}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}\right]^2\right] \\
&\leq \exp\left[\frac{2|\mathbf{j}^{\max}|}{B^2}\right] \mathop{\mathbf{E}}_{S_{\bar{\mathbf{j}}}\sim D^{m-|\mathbf{j}\cup\mathbf{j}'|}} \exp\left[2(m-|\mathbf{j}\cup\mathbf{j}'|)\left[\left(\frac{1}{2} - \frac{\mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}}{2B}\right) - \left(\frac{1}{2} - \frac{\mathcal{M}^{D}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}}{2B}\right)\right]^2\right].
\end{aligned}$$

We know $2(a-b)^2 \leq \mathrm{kl}(a\|b)$ is valid for any $a, b \in [0,1]$ provided that if $a = 0$ then so is b and if $a = 1$ then so is b. Since the elements of $\mathcal{H}^S$ are B-bounded and $S_{\bar{\mathbf{j}}}$ is *i.i.d.* from D, we have:

$$\mathcal{M}^{D}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}} = -B^2 \Rightarrow \mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}} = -B^2, \quad \text{and} \quad \mathcal{M}^{D}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}} = B^2 \Rightarrow \mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}} = B^2.$$

Then: $\frac{1}{2} - \frac{\mathcal{M}^{D}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}}{2B^2} = 0 \Rightarrow \frac{1}{2} - \frac{\mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}}{2B^2} = 0$, and $\frac{1}{2} - \frac{\mathcal{M}^{D}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}}{2B^2} = 1 \Rightarrow \frac{1}{2} - \frac{\mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}}{2B^2} = 1$.

Since: $0 \leq \frac{1}{2} - \frac{\mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}}{2B^2} \leq 1$, and $0 \leq \frac{1}{2} - \frac{\mathcal{M}^{D}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}}{2B^2} \leq 1$, we have:

$$\mathop{\mathbf{E}}_{S\sim D^m} X_P \leq \exp\left[\frac{2|\mathbf{j}^{\max}|}{B^2}\right] \mathop{\mathbf{E}}_{S_{\bar{\mathbf{j}}}\sim D^{m-|\mathbf{j}\cup\mathbf{j}'|}} \exp\left[(m-|\mathbf{j}\cup\mathbf{j}'|)\,\mathrm{kl}\left(\tfrac{1}{2} - \frac{\mathcal{M}^{S_{\bar{\mathbf{j}}}}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}}{2B^2}\middle\|\tfrac{1}{2} - \frac{\mathcal{M}^{D}_{h^{\omega}_{S_{\mathbf{j}}},h^{\omega'}_{S_{\mathbf{j}'}}}}{2B^2}\right)\right].$$

By applying Maurer's Lemma (Lemma 1), we obtain:

$$\begin{aligned}
\mathop{\mathbf{E}}_{S\sim D^m} X_P &\leq \exp\left(\frac{2|\mathbf{j}^{\max}|}{B^2}\right) \mathop{\mathbf{E}}_{S_{\bar{\mathbf{j}}}\sim D^{m-|\mathbf{j}\cup\mathbf{j}'|}} 2\sqrt{(m-|\mathbf{j}\cup\mathbf{j}'|)} \leq \exp\left(\frac{2|\mathbf{j}^{\max}|}{B^2}\right) 2\sqrt{(m-|\mathbf{j}\cup\mathbf{j}'|)} \\
&\leq \exp\left(\frac{2|\mathbf{j}^{\max}|}{B^2}\right) 2\sqrt{m}.
\end{aligned}$$

Finally: $\mathop{\mathbf{Pr}}_{S\sim D^m}\left(\begin{array}{l}\text{for all } \mathbf{P}\text{-aligned distribution } Q \text{ on } \mathcal{H}^S, \\ |\mathcal{M}^{D}_{Q^2} - \mathcal{M}^{S}_{Q^2}| \leq \dfrac{2B^2\sqrt{\dfrac{2|\mathbf{j}^{\max}|}{B^2} + \ln\left(\dfrac{2\sqrt{m}}{\delta}\right)}}{\sqrt{2(m-2|\mathbf{j}^{\max}|)}}\end{array}\right) \geq 1-\delta$