Chapter 3

Transductive PAC-Bayesian learning

3.1. Basic inequalities

3.1.1. The transductive setting. In this chapter the *observed* sample $(X_i, Y_i)_{i=1}^N$ will be supplemented with a *test* or *shadow* sample $(X_i, Y_i)_{i=N+1}^{(k+1)N}$. This point of view, called *transductive classification*, has been introduced by V. Vapnik. It may be justified in different ways.

On the practical side, one interest of the transductive setting is that it is often a lot easier to collect examples than it is to label them, so that it is not unrealistic to assume that we indeed have two training samples, one labelled and one unlabelled. It also covers the case when a batch of patterns is to be classified and we are allowed to observe the whole batch before issuing the classification.

On the mathematical side, considering a shadow sample proves technically fruitful. Indeed, when introducing the Vapnik–Cervonenkis entropy and Vapnik–Cervonenkis dimension concepts, as well as when dealing with compression schemes, albeit the *inductive* setting is our final concern, the transductive setting is a useful detour. In this second scenario, intermediate technical results involving the shadow sample are integrated with respect to unobserved random variables in a second stage of the proofs.

Let us describe now the changes to be made to previous notation to adapt them to the transductive setting. The distribution $\mathbb{P}$ will be a probability measure on the canonical space $\Omega = (\mathcal{X} \times \mathcal{Y})^{(k+1)N}$, and $(X_i, Y_i)_{i=1}^{(k+1)N}$ will be the canonical process on this space (that is the coordinate process). Unless explicitly mentioned, the parameter k indicating the size of the shadow sample will remain fixed. Assuming the shadow sample size is a multiple of the training sample size is convenient without significantly restricting generality. For a while, we will use a weaker assumption than independence, assuming that $\mathbb{P}$ is *partially exchangeable*, since this is all we need in the proofs.

Definition 3.1.1. For $i = 1, \ldots, N$, let $\tau_i : \Omega \to \Omega$ be defined for any

$\omega = (\omega_j)_{j=1}^{(k+1)N} \in \Omega$ by

$$\begin{cases} \tau_i(\omega)_{i+jN} = \omega_{i+(j-1)N}, & j = 1, \ldots, k, \\ \tau_i(\omega)_i = \omega_{i+kN}, & \\ \text{and } \tau_i(\omega)_{m+jN} = \omega_{m+jN}, & m \neq i, m = 1, \ldots, N, j = 0, \ldots k. \end{cases}$$

Clearly, if we arrange the $(k+1)N$ samples in a $N \times (k+1)$ array, τ_i performs a circular permutation of $k+1$ entries on the ith row, leaving the other rows unchanged. Moreover, all the circular permutations of the ith row have the form τ_i^j, j ranging from 0 to k.

The probability distribution $\mathbb{P}$ is said to be partially exchangeable if for any $i = 1, \ldots, N$, $\mathbb{P} \circ \tau_i^{-1} = \mathbb{P}$.

This means equivalently that for any bounded measurable function $h : \Omega \to \mathbb{R}$, $\mathbb{P}(h \circ \tau_i) = \mathbb{P}(h)$.

In the same way a function h defined on Ω will be said to be partially exchangeable if $h \circ \tau_i = h$ for any $i = 1, \ldots, N$. Accordingly a posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta, \mathcal{T})$ will be said to be partially exchangeable when $\rho(\omega, A) = \rho\big[\tau_i(\omega), A\big]$, for any $\omega \in \Omega$, any $i = 1, \ldots, N$ and any $A \in \mathcal{T}$.

For any bounded measurable function h, let us define $T_i(h) = \frac{1}{k+1} \sum_{j=0}^{k} h \circ \tau_i^j$. Let $T(h) = T_N \circ \cdots \circ T_1(h)$. For any partially exchangeable probability distribution $\mathbb{P}$, and for any bounded measurable function h, $\mathbb{P}\big[T(h)\big] = \mathbb{P}(h)$. Let us put

$$\sigma_i(\theta) = \mathbb{1}\big[f_\theta(X_i) \neq Y_i\big], \quad \text{indicating the success or failure of } f_\theta \text{ to predict } Y_i \text{ from } X_i,$$

$$r_1(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sigma_i(\theta), \quad \text{the empirical error rate of } f_\theta \text{ on the observed sample,}$$

$$r_2(\theta) = \frac{1}{kN} \sum_{i=N+1}^{(k+1)N} \sigma_i(\theta), \quad \text{the error rate of } f_\theta \text{ on the shadow sample,}$$

$$\overline{r}(\theta) = \frac{r_1(\theta) + k r_2(\theta)}{k+1} = \frac{1}{(k+1)N} \sum_{i=1}^{(k+1)N} \sigma_i(\theta), \quad \text{the global error rate of } f_\theta,$$

$$R_i(\theta) = \mathbb{P}\big[f_\theta(X_i) \neq Y_i\big], \quad \text{the expected error rate of } f_\theta \text{ on the } i\text{th input,}$$

$$R(\theta) = \frac{1}{N} \sum_{i=1}^{N} R_i(\theta) = \mathbb{P}\big[r_1(\theta)\big] = \mathbb{P}\big[r_2(\theta)\big], \quad \text{the average expected error rate of } f_\theta \text{ on all inputs.}$$

We will allow for posterior distributions $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$ depending on the shadow sample. The most interesting ones will anyhow be independent of the shadow labels $Y_{N+1}, \ldots, Y_{(k+1)N}$. We will be interested in the conditional expected error rate of the randomized classification rule described by ρ on the shadow sample, given the observed sample, that is, $\mathbb{P}\big[\rho(r_2) | (X_i, Y_i)_{i=1}^N\big]$. This is a natural extension of the notion of *generalization error rate*: this is indeed the error rate to be expected when the randomized classification rule described by the posterior distribution ρ is applied to the shadow sample (which should in this case more purposefully be called the test sample).

To see the connection with the previously defined generalization error rate, let us comment on the case when $\mathbb{P}$ is invariant by any permutation of any row, meaning that

$$\mathbb{P}\big[h(\omega \circ s)\big] = \mathbb{P}\big[h(\omega)\big] \text{ for all } s \in \mathfrak{S}(\{i + jN; j = 0, \dots, k\})$$

and all $i = 1, \dots, N$, where $\mathfrak{S}(A)$ is the set of permutations of A, extended to $\{1, \dots, (k+1)N\}$ so as to be the identity outside of A. In other words, $\mathbb{P}$ is assumed to be invariant under any permutation which keeps the rows unchanged. In this case, if ρ is invariant by any permutation of any row of the shadow sample, meaning that $\rho(\omega \circ s) = \rho(\omega) \in \mathcal{M}_+^1(\Theta)$, $s \in \mathfrak{S}(\{i + jN; j = 1, \dots, k\})$, $i = 1, \dots, N$, then $\mathbb{P}\big[\rho(r_2)|(X_i, Y_i)_{i=1}^N\big] = \frac{1}{N}\sum_{i=1}^N \mathbb{P}\big[\rho(\sigma_{i+N})|(X_i, Y_i)_{i=1}^N\big]$, meaning that the expectation can be taken on a restricted shadow sample of the same size as the observed sample. If moreover the rows are equidistributed, meaning that their marginal distributions are equal, then

$$\mathbb{P}\big[\rho(r_2)|(X_i, Y_i)_{i=1}^N\big] = \mathbb{P}\big[\rho(\sigma_{N+1})|(X_i, Y_i)_{i=1}^N\big].$$

This means that under these quite commonly fulfilled assumptions, the expectation can be taken on a single new object to be classified, our study thus covers the case when only one of the patterns from the shadow sample is to be labelled and one is interested in the expected error rate of this single labelling. Of course, in the case when $\mathbb{P}$ is i.i.d. and ρ depends only on the training sample $(X_i, Y_i)_{i=1}^N$, we fall back on the usual criterion of performance $\mathbb{P}\big[\rho(r_2)|(Z_i)_{i=1}^N\big] = \rho(R) = \rho(R_1)$.

3.1.2. Absolute bound. Using an obvious factorization, and considering for the moment a fixed value of θ and any partially exchangeable positive real measurable function $\lambda : \Omega \to \mathbb{R}_+$, we can compute the log-Laplace transform of r_1 under T, which acts like a conditional probability distribution:

$$\log\Big\{T\big[\exp(-\lambda r_1)\big]\Big\} = \sum_{i=1}^N \log\Big\{T_i\big[\exp(-\tfrac{\lambda}{N}\sigma_i)\big]\Big\}$$
$$\leq N \log\left\{\frac{1}{N}\sum_{i=1}^N T_i\Big[\exp\big(-\tfrac{\lambda}{N}\sigma_i\big)\Big]\right\} = -\lambda\Phi_{\frac{\lambda}{N}}(\overline{r}),$$

where the function $\Phi_{\frac{\lambda}{N}}$ was defined by equation (1.1, page 2). Remarking that $T\Big\{\exp\Big[\lambda\big[\Phi_{\frac{\lambda}{N}}(\overline{r}) - r_1\big]\Big]\Big\} = \exp\big[\lambda\Phi_{\frac{\lambda}{N}}(\overline{r})\big]T\big[\exp(-\lambda r_1)\big]$ we obtain

Lemma 3.1.1. *For any $\theta \in \Theta$ and any partially exchangeable positive real measurable function $\lambda : \Omega \to \mathbb{R}_+$,*

$$T\Big\{\exp\Big[\lambda\big\{\Phi_{\frac{\lambda}{N}}\big[\overline{r}(\theta)\big] - r_1(\theta)\big\}\Big]\Big\} \leq 1.$$

We deduce from this lemma a result analogous to the inductive case:

Theorem 3.1.2. *For any partially exchangeable positive real measurable function $\lambda : \Omega \times \Theta \to \mathbb{R}_+$, for any partially exchangeable posterior distribution $\pi : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$\mathbb{P}\left\{\exp\left[\sup_{\rho \in \mathcal{M}_+^1(\Theta)} \rho\Big[\lambda\big[\Phi_{\frac{\lambda}{N}}(\overline{r}) - r_1\big]\Big] - \mathcal{K}(\rho, \pi)\right]\right\} \leq 1.$$

The proof is deduced from the previous lemma, using the fact that π is partially exchangeable:

$$\mathbb{P}\biggl\{\exp\biggl[\sup_{\rho\in\mathcal{M}_+^1(\Theta)}\rho\Bigl[\lambda\bigl[\Phi_{\frac{\lambda}{N}}(\overline{r})-r_1\bigr]\Bigr]-\mathcal{K}(\rho,\pi)\biggr]\biggr\}$$
$$=\mathbb{P}\biggl\{\pi\Bigl\{\exp\Bigl[\lambda\bigl[\Phi_{\frac{\lambda}{N}}(\overline{r})-r_1\bigr]\Bigr]\Bigr\}\biggr\}=\mathbb{P}\biggl\{T\pi\Bigl\{\exp\Bigl[\lambda\bigl[\Phi_{\frac{\lambda}{N}}(\overline{r})-r_1\bigr]\Bigr]\Bigr\}\biggr\}$$
$$=\mathbb{P}\biggl\{\pi\Bigl\{T\exp\Bigl[\lambda\bigl[\Phi_{\frac{\lambda}{N}}(\overline{r})-r_1\bigr]\Bigr]\Bigr\}\biggr\}\leq 1.$$

3.1.3. Relative bounds. Introducing in the same way

$$m'(\theta,\theta')=\frac{1}{N}\sum_{i=1}^{N}\Bigl|\mathbb{1}\bigl[f_\theta(X_i)\neq Y_i\bigr]-\mathbb{1}\bigl[f_{\theta'}(X_i)\neq Y_i\bigr]\Bigr|$$
$$\text{and}\quad \overline{m}(\theta,\theta')=\frac{1}{(k+1)N}\sum_{i=1}^{(k+1)N}\Bigl|\mathbb{1}\bigl[f_\theta(X_i)\neq Y_i\bigr]-\mathbb{1}\bigl[f_{\theta'}(X_i)\neq Y_i\bigr]\Bigr|,$$

we could prove along the same line of reasoning

Theorem 3.1.3. *For any real parameter λ, any $\widetilde{\theta}\in\Theta$, any partially exchangeable posterior distribution $\pi:\Omega\to\mathcal{M}_+^1(\Theta)$,*

$$\mathbb{P}\biggl\{\exp\biggl[\sup_{\rho\in\mathcal{M}_+^1(\Theta)}\lambda\Bigl[\rho\bigl\{\Psi_{\frac{\lambda}{N}}\bigl[\overline{r}(\cdot)-\overline{r}(\widetilde{\theta}),\overline{m}(\cdot,\widetilde{\theta})\bigr]\bigr\}$$
$$-\bigl[\rho(r_1)-r_1(\widetilde{\theta})\bigr]\Bigr]-\mathcal{K}(\rho,\pi)\biggr]\biggr\}\leq 1,$$

where the function $\Psi_{\frac{\lambda}{N}}$ was defined by equation (1.21, page 35).

Theorem 3.1.4. *For any real constant γ, for any $\widetilde{\theta}\in\Theta$, for any partially exchangeable posterior distribution $\pi:\Omega\to\mathcal{M}_+^1(\Theta)$,*

$$\mathbb{P}\biggl\{\exp\biggl[\sup_{\rho\in\mathcal{M}_+^1(\Theta)}\biggl\{-N\rho\Bigl\{\log\Bigl[1-\tanh\bigl(\tfrac{\gamma}{N}\bigr)\bigl[\overline{r}(\cdot)-\overline{r}(\widetilde{\theta})\bigr]\Bigr]\Bigr\}$$
$$-\gamma\bigl[\rho(r_1)-r_1(\widetilde{\theta})\bigr]-N\log\bigl[\cosh\bigl(\tfrac{\gamma}{N}\bigr)\bigr]\rho\bigl[m'(\cdot,\widetilde{\theta})\bigr]-\mathcal{K}(\rho,\pi)\biggr\}\biggr]\biggr\}\leq 1.$$

This last theorem can be generalized to give

Theorem 3.1.5. *For any real constant γ, for any partially exchangeable posterior distributions $\pi^1,\pi^2:\Omega\to\mathcal{M}_+^1(\Theta)$,*

$$\mathbb{P}\biggl\{\exp\biggl[\sup_{\rho_1,\rho_2\in\mathcal{M}_+^1(\Theta)}\biggl\{-N\log\Bigl\{1-\tanh\bigl(\tfrac{\gamma}{N}\bigr)\bigl[\rho_1(\overline{r})-\rho_2(\overline{r})\bigr]\Bigr\}$$
$$-\gamma\bigl[\rho_1(r_1)-\rho_2(r_1)\bigr]-N\log\bigl[\cosh\bigl(\tfrac{\gamma}{N}\bigr)\bigr]\rho_1\otimes\rho_2(m')$$
$$-\mathcal{K}(\rho_1,\pi^1)-\mathcal{K}(\rho_2,\pi^2)\biggr\}\biggr]\biggr\}\leq 1.$$

To conclude this section, we see that the basic theorems of transductive PAC-Bayesian classification have exactly the same form as the basic inequalities of inductive classification, Theorems 1.1.4 (page 4), 1.4.2 (page 35) and 1.4.3 (page 37) *with* $R(\theta)$ *replaced with* $\overline{r}(\theta)$, $r(\theta)$ replaced with $r_1(\theta)$ and $M'(\theta, \widetilde{\theta})$ replaced with $\overline{m}(\theta, \widetilde{\theta})$.

Thus all the results of the first two chapters remain true under the hypotheses of transductive classification, with $R(\theta)$ *replaced with* $\overline{r}(\theta)$, $r(\theta)$ *replaced with* $r_1(\theta)$ *and* $M'(\theta, \widetilde{\theta})$ *replaced with* $\overline{m}(\theta, \widetilde{\theta})$.

Consequently, in the case when the unlabelled shadow sample is observed, it is possible to improve on the Vapnik bounds to be discussed hereafter by using an explicit partially exchangeable posterior distribution π *and resorting to localized or to relative bounds (in the case at least of unlimited computing resources, which of course may still be unrealistic in many real world situations, and with the caveat, to be recalled in the conclusion of this study, that for small sample sizes and comparatively complex classification models, the improvement may not be so decisive).*

Let us notice also that the transductive setting when experimentally available, has the advantage that

$$\overline{d}(\theta, \theta') = \frac{1}{(k+1)N} \sum_{i=1}^{(k+1)N} \mathbb{1}\big[f_{\theta'}(X_i) \neq f_\theta(X_i)\big] \geq \overline{m}(\theta, \theta') \geq \overline{r}(\theta) - \overline{r}(\theta'), \qquad \theta, \theta' \in \Theta,$$

is observable in this context, providing an empirical upper bound for the difference $\overline{r}(\widehat{\theta}) - \rho(\overline{r})$ for any non-randomized estimator $\widehat{\theta}$ and any posterior distribution ρ, namely

$$\overline{r}(\widehat{\theta}) \leq \rho(\overline{r}) + \rho\big[\,\overline{d}(\cdot, \widehat{\theta})\big].$$

Thus in the setting of transductive statistical experiments, the PAC-Bayesian framework provides fully empirical bounds for the error rate of non-randomized estimators $\widehat{\theta} : \Omega \to \Theta$, even when using a non-atomic prior π (or more generally a non-atomic partially exchangeable posterior distribution π), even when Θ is not a vector space and even when $\theta \mapsto R(\theta)$ cannot be proved to be convex on the support of some useful posterior distribution ρ.

3.2. Vapnik bounds for transductive classification

In this section, we will stick to plain unlocalized non-relative bounds. As we have already mentioned, (and as it was put forward by Vapnik himself in his seminal works), these bounds are not always superseded by the asymptotically better ones when the sample is of small size: they deserve all our attention for this reason. We will start with the general case of a shadow sample of arbitrary size. We will then discuss the case of a shadow sample of equal size to the training set and the case of a fully exchangeable sample distribution, showing how they can be taken advantage of to sharpen inequalities.

3.2.1. With a shadow sample of arbitrary size. The great thing with the transductive setting is that we are manipulating only r_1 and $\overline{r}$ which can take only a finite number of values and therefore are piecewise constant on Θ. This makes it possible to derive inequalities that will hold uniformly for any value of the parameter

$\theta \in \Theta$. To this purpose, let us consider for any value $\theta \in \Theta$ of the parameter the subset $\Delta(\theta) \subset \Theta$ of parameters θ' such that the classification rule $f_{\theta'}$ answers the same on the extended sample $(X_i)_{i=1}^{(k+1)N}$ as f_θ. Namely, let us put for any $\theta \in \Theta$

$$\Delta(\theta) = \big\{\theta' \in \Theta; f_{\theta'}(X_i) = f_\theta(X_i), i = 1, \dots, (k+1)N\big\}.$$

We see immediately that $\Delta(\theta)$ is an exchangeable parameter subset on which r_1 and r_2 and therefore also $\overline{r}$ take constant values. Thus for any $\theta \in \Theta$ we may consider the posterior ρ_θ defined by

$$\frac{d\rho_\theta}{d\pi}(\theta') = \mathbb{1}\big[\theta' \in \Delta(\theta)\big]\pi\big[\Delta(\theta)\big]^{-1},$$

and use the fact that $\rho_\theta(r_1) = r_1(\theta)$ and $\rho_\theta(\overline{r}) = \overline{r}(\theta)$, to prove that

LEMMA 3.2.1. *For any partially exchangeable positive real measurable function $\lambda : \Omega \times \Theta \to \mathbb{R}$ such that*

$$\lambda(\omega, \theta') = \lambda(\omega, \theta), \quad \theta \in \Theta, \theta' \in \Delta(\theta), \omega \in \Omega, \tag{3.1}$$

and any partially exchangeable posterior distribution $\pi : \Omega \to \mathcal{M}_+^1(\Theta)$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$,

$$\Phi_{\frac{\lambda}{N}}\big[\overline{r}(\theta)\big] + \frac{\log\big\{\epsilon\pi\big[\Delta(\theta)\big]\big\}}{\lambda(\theta)} \le r_1(\theta).$$

We can then remark that for any value of λ independent of ω, the left-hand side of the previous inequality is a partially exchangeable function of $\omega \in \Omega$. Thus this left-hand side is maximized by some partially exchangeable function λ, namely

$$\arg\max_\lambda \left\{ \Phi_{\frac{\lambda}{N}}\big[\overline{r}(\theta)\big] + \frac{\log\big\{\epsilon\pi\big[\Delta(\theta)\big]\big\}}{\lambda} \right\}$$

is partially exchangeable as depending only on partially exchangeable quantities. Moreover this choice of $\lambda(\omega, \theta)$ satisfies also condition (3.1) stated in the previous lemma of being constant on $\Delta(\theta)$, proving

LEMMA 3.2.2. *For any partially exchangeable posterior distribution $\pi : \Omega \to \mathcal{M}_+^1(\Theta)$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$ and any $\lambda \in \mathbb{R}_+$,*

$$\Phi_{\frac{\lambda}{N}}\big[\overline{r}(\theta)\big] + \frac{\log\big\{\epsilon\pi\big[\Delta(\theta)\big]\big\}}{\lambda} \le r_1(\theta).$$

Writing $\overline{r} = \frac{r_1 + kr_2}{k+1}$ and rearranging terms we obtain

THEOREM 3.2.3. *For any partially exchangeable posterior distribution $\pi : \Omega \to \mathcal{M}_+^1(\Theta)$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$,*

$$r_2(\theta) \le \frac{k+1}{k} \inf_{\lambda \in \mathbb{R}_+} \frac{1 - \exp\left(-\frac{\lambda}{N} r_1(\theta) + \frac{\log\big\{\epsilon\pi\big[\Delta(\theta)\big]\big\}}{N}\right)}{1 - \exp\big(-\frac{\lambda}{N}\big)} - \frac{r_1(\theta)}{k}.$$

If we have a set of binary classification rules $\{f_\theta; \theta \in \Theta\}$ whose Vapnik–Cervonenkis dimension is not greater than h, we can choose π such that $\pi\big[\Delta(\theta)\big]$ is independent of θ and not less than $\left(\frac{h}{e(k+1)N}\right)^h$, as will be proved further on in Theorem 4.2.2 (page 144).

Another important setting where the complexity term $-\log\big\{\pi\big[\Delta(\theta)\big]\big\}$ can easily be controlled is the case of *compression schemes*, introduced by Little et al. (1986). It goes as follows: we are given for each labelled sub-sample $(X_i, Y_i)_{i\in J}$, $J \subset \{1, \ldots, N\}$, an estimator of the parameter

$$\widehat{\theta}\big[(X_i, Y_i)_{i\in J}\big] = \widehat{\theta}_J, \quad J \subset \{1, \ldots, N\}, |J| \le h,$$

where

$$\widehat{\theta} : \bigsqcup_{k=1}^{N} \big(\mathcal{X} \times \mathcal{Y}\big)^k \to \Theta$$

is an exchangeable function providing estimators for sub-samples of arbitrary size. Let us assume that $\widehat{\theta}$ is exchangeable, meaning that for any $k = 1, \ldots, N$ and any permutation σ of $\{1, \ldots, k\}$

$$\widehat{\theta}\big[(x_i, y_i)_{i=1}^k\big] = \widehat{\theta}\big[(x_{\sigma(i)}, y_{\sigma(i)})_{i=1}^k\big], \qquad (x_i, y_i)_{i=1}^k \in \big(\mathcal{X} \times \mathcal{Y}\big)^k.$$

In this situation, we can introduce the exchangeable subset

$$\Big\{\widehat{\theta}_J; J \subset \{1, \ldots, (k+1)N\}, |J| \le h\Big\} \subset \Theta,$$

which is seen to contain at most

$$\sum_{j=0}^{h} \binom{(k+1)N}{j} \le \left(\frac{e(k+1)N}{h}\right)^h$$

classification rules — as will be proved later on in Theorem 4.2.3 (page 144). Note that we had to extend the range of J to all the subsets of the extended sample, although we will use for estimation only those of the training sample, on which the labels are observed. Thus in this case also we can find a partially exchangeable posterior distribution π such that

$$\pi\big[\Delta(\widehat{\theta}_J)\big] \ge \left(\frac{h}{e(k+1)N}\right)^h.$$

We see that the size of the compression scheme plays the same role in this complexity bound as the Vapnik–Cervonenkis dimension for Vapnik–Cervonenkis classes.

In these two cases of binary classification with Vapnik–Cervonenkis dimension not greater than h and compression schemes depending on a compression set with at most h points, we get a bound of

$$r_2(\theta) \le \frac{k+1}{k} \inf_{\lambda \in \mathbb{R}_+} \frac{1 - \exp\left(-\frac{\lambda}{N} r_1(\theta) - \frac{h \log\left(\frac{e(k+1)N}{h}\right) - \log(\epsilon)}{N}\right)}{1 - \exp\big(-\frac{\lambda}{N}\big)} - \frac{r_1(\theta)}{k}.$$

Let us make some numerical application: when $N = 1000, h = 10, \epsilon = 0.01$, and $\inf_\Theta r_1 = r_1(\widehat{\theta}) = 0.2$, we find that $r_2(\widehat{\theta}) \leq 0.4093$, for k between 15 and 17, and values of λ equal respectively to 965, 968 and 971. For $k = 1$, we find only $r_2(\widehat{\theta}) \leq 0.539$, showing the interest of allowing k to be larger than 1.

3.2.2. When the shadow sample has the same size as the training sample. In the case when $k = 1$, we can improve Theorem 3.1.2 by taking advantage of the fact that $T_i(\sigma_i)$ can take only 3 values, namely 0, 0.5 and 1. We see thus that $T_i(\sigma_i) - \Phi_{\frac{\lambda}{N}}\big[T_i(\sigma_i)\big]$ can take only two values, 0 and $\frac{1}{2} - \Phi_{\frac{\lambda}{N}}(\frac{1}{2})$, because $\Phi_{\frac{\lambda}{N}}(0) = 0$ and $\Phi_{\frac{\lambda}{N}}(1) = 1$. Thus

$$T_i(\sigma_i) - \Phi_{\frac{\lambda}{N}}\big[T_i(\sigma_i)\big] = \big[1 - |1 - 2T_i(\sigma_i)|\big]\big[\tfrac{1}{2} - \Phi_{\frac{\lambda}{N}}(\tfrac{1}{2})\big].$$

This shows that in the case when $k = 1$,

$$\begin{aligned}\log\Big\{T\big[\exp(-\lambda r_1)\big]\Big\} &= -\lambda\overline{r} + \frac{\lambda}{N}\sum_{i=1}^N T_i(\sigma_i) - \Phi_{\frac{\lambda}{N}}\big[T_i(\sigma_i)\big] \\ &= -\lambda\overline{r} + \frac{\lambda}{N}\sum_{i=1}^N \big[1 - |1 - 2T_i(\sigma_i)|\big]\big[\tfrac{1}{2} - \Phi_{\frac{\lambda}{N}}(\tfrac{1}{2})\big] \\ &\leq -\lambda\overline{r} + \lambda\big[\tfrac{1}{2} - \Phi_{\frac{\lambda}{N}}(\tfrac{1}{2})\big]\big[1 - |1 - 2\overline{r}|\big].\end{aligned}$$

Noticing that $\frac{1}{2} - \Phi_{\frac{\lambda}{N}}(\frac{1}{2}) = \frac{N}{\lambda}\log\big[\cosh(\frac{\lambda}{2N})\big]$, we obtain

Theorem 3.2.4. *For any partially exchangeable function $\lambda : \Omega \times \Theta \to \mathbb{R}_+$, for any partially exchangeable posterior distribution $\pi : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$\mathbb{P}\Bigg\{\exp\Bigg[\sup_{\rho\in\mathcal{M}_+^1(\Theta)} \rho\Big[\lambda(\overline{r} - r_1) - N\log\big[\cosh(\tfrac{\lambda}{2N})\big]\big(1 - |1 - 2\overline{r}|\big)\Big] - \mathcal{K}(\rho,\pi)\Bigg]\Bigg\} \leq 1.$$

As a consequence, reasoning as previously, we deduce

Theorem 3.2.5. *In the case when $k = 1$, for any partially exchangeable posterior distribution $\pi : \Omega \to \mathcal{M}_+^1(\Theta)$, with $\mathbb{P}$ probability at least $1 - \epsilon$, for any $\theta \in \Theta$ and any $\lambda \in \mathbb{R}_+$,*

$$\overline{r}(\theta) - \tfrac{N}{\lambda}\log\big[\cosh(\tfrac{\lambda}{2N})\big]\big(1 - |1 - 2\overline{r}(\theta)|\big) + \frac{\log\big\{\epsilon\pi\big[\Delta(\theta)\big]\big\}}{\lambda} \leq r_1(\theta);$$

and consequently for any $\theta \in \Theta$,

$$r_2(\theta) \leq 2\inf_{\lambda\in\mathbb{R}_+} \frac{r_1(\theta) - \dfrac{\log\big\{\epsilon\pi\big[\Delta(\theta)\big]\big\}}{\lambda}}{1 - \frac{2N}{\lambda}\log\big[\cosh(\frac{\lambda}{2N})\big]} - r_1(\theta).$$

In the case of binary classification using a Vapnik–Cervonenkis class of Vapnik–Cervonenkis dimension not greater than h, we can choose π such that $-\log\big\{\pi\big[\Delta(\theta)\big]\big\} \leq h\log(\frac{2eN}{h})$ and obtain the following numerical illustration of

this theorem: for $N = 1000$, $h = 10$, $\epsilon = 0.01$ and $\inf_\Theta r_1 = r_1(\widehat{\theta}) = 0.2$, we find an upper bound $r_2(\widehat{\theta}) \leq 0.5033$, which improves on Theorem 3.2.3 but still is not under the significance level $\frac{1}{2}$ (achieved by blind random classification). This indicates that considering shadow samples of arbitrary sizes some noisy situations yields a significant improvement on bounds obtained with a shadow sample of the same size as the training sample.

3.2.3. When moreover the distribution of the augmented sample is exchangeable. When $k = 1$ and $\mathbb{P}$ is exchangeable meaning that for any bounded measurable function $h : \Omega \to \mathbb{R}$ and any permutation $s \in \mathfrak{S}\big(\{1, \ldots, 2N\}\big)$ $\mathbb{P}\big[h(\omega \circ s)\big] = \mathbb{P}\big[h(\omega)\big]$, then we can still improve the bound as follows. Let

$$T'(h) = \frac{1}{N!} \sum_{s \in \mathfrak{S}\left(\{N+1,\ldots,2N\}\right)} h(\omega \circ s).$$

Then we can write

$$1 - |1 - 2T_i(\sigma_i)| = (\sigma_i - \sigma_{i+N})^2 = \sigma_i + \sigma_{i+N} - 2\sigma_i\sigma_{i+N}.$$

Using this identity, we get for any exchangeable function $\lambda : \Omega \times \Theta \to \mathbb{R}_+$,

$$T\Big\{\exp\Big[\lambda(\overline{r} - r_1) - \log\big[\cosh(\tfrac{\lambda}{2N})\big] \sum_{i=1}^{N} \big(\sigma_i + \sigma_{i+N} - 2\sigma_i\sigma_{i+N}\big)\Big]\Big\} \leq 1.$$

Let us put

$$A(\lambda) = \tfrac{2N}{\lambda} \log\big[\cosh(\tfrac{\lambda}{2N})\big], \tag{3.2}$$

$$v(\theta) = \frac{1}{2N} \sum_{i=1}^{N} (\sigma_i + \sigma_{i+N} - 2\sigma_i\sigma_{i+N}). \tag{3.3}$$

With this notation

$$T\Big\{\exp\big\{\lambda\big[\overline{r} - r_1 - A(\lambda)v\big]\big\}\Big\} \leq 1.$$

Let us notice now that

$$T'\big[v(\theta)\big] = \overline{r}(\theta) - r_1(\theta)r_2(\theta).$$

Let $\pi : \Omega \to \mathcal{M}_+^1(\Theta)$ be any given exchangeable posterior distribution. Using the exchangeability of $\mathbb{P}$ and π and the exchangeability of the exponential function, we get

$$\begin{aligned}
&\mathbb{P}\Big\{\pi\Big[\exp\big\{\lambda\big[\overline{r} - r_1 - A(\overline{r} - r_1 r_2)\big]\big\}\Big]\Big\} = \mathbb{P}\Big\{\pi\Big[\exp\big\{\lambda\big[\overline{r} - r_1 - AT'(v)\big]\big\}\Big]\Big\} \\
&\quad \leq \mathbb{P}\Big\{\pi\Big[T' \exp\big\{\lambda\big[\overline{r} - r_1 - Av\big]\big\}\Big]\Big\} = \mathbb{P}\Big\{T'\pi\Big[\exp\big\{\lambda\big[\overline{r} - r_1 - Av\big]\big\}\Big]\Big\} \\
&\quad = \mathbb{P}\Big\{\pi\Big[\exp\big\{\lambda\big[\overline{r} - r_1 - Av\big]\big\}\Big]\Big\} = \mathbb{P}\Big\{T\pi\Big[\exp\big\{\lambda\big[\overline{r} - r_1 - Av\big]\big\}\Big]\Big\} \\
&\quad = \mathbb{P}\Big\{\pi\Big[T \exp\big\{\lambda\big[\overline{r} - r_1 - Av\big]\big\}\Big]\Big\} \leq 1.
\end{aligned}$$

We are thus ready to state

THEOREM 3.2.6. *In the case when $k = 1$, for any exchangeable probability distribution $\mathbb{P}$, for any exchangeable posterior distribution $\pi : \Omega \to \mathcal{M}_+^1(\Theta)$, for any exchangeable function $\lambda : \Omega \times \Theta \to \mathbb{R}_+$,*

$$\mathbb{P}\biggl\{\exp\biggl[\sup_{\rho\in\mathcal{M}_+^1(\Theta)} \rho\Bigl\{\lambda\bigl[\overline{r} - r_1 - A(\lambda)(\overline{r} - r_1 r_2)\bigr]\Bigr\} - \mathcal{K}(\rho,\pi)\biggr]\biggr\} \leq 1,$$

where $A(\lambda)$ is defined by equation (3.2, page 119).

We then deduce as previously

COROLLARY 3.2.7. *For any exchangeable posterior distribution $\pi : \Omega \to \mathcal{M}_+^1(\Theta)$, for any exchangeable probability measure $\mathbb{P} \in \mathcal{M}_+^1(\Omega)$, for any measurable exchangeable function $\lambda : \Omega \times \Theta \to \mathbb{R}_+$, with $\mathbb{P}$ probability at least $1 - \epsilon$, for any $\theta \in \Theta$,*

$$\overline{r}(\theta) \leq r_1(\theta) + A(\lambda)\bigl[\overline{r}(\theta) - r_1(\theta) r_2(\theta)\bigr] - \frac{\log\bigl\{\epsilon\pi\bigl[\Delta(\theta)\bigr]\bigr\}}{\lambda},$$

where $A(\lambda)$ is defined by equation (3.2, page 119).

In order to deduce an empirical bound from this theorem, we have to make some choice for $\lambda(\omega,\theta)$. Fortunately, it is easy to show that the bound holds uniformly in λ, because the inequality can be rewritten as a function of only one non-exchangeable quantity, namely $r_1(\theta)$. Indeed, since $r_2 = 2\overline{r} - r_1$, we see that the inequality can be written as

$$\overline{r}(\theta) \leq r_1(\theta) + A(\lambda)\bigl[\overline{r}(\theta) - 2\overline{r}(\theta) r_1(\theta) + r_1(\theta)^2\bigr] - \frac{\log\bigl\{\epsilon\pi\bigl[\Delta(\theta)\bigr]}{\lambda}.$$

It can be solved in $r_1(\theta)$, to get

$$r_1(\theta) \geq f\Bigl(\lambda, \overline{r}(\theta), -\log\bigl\{\epsilon\pi\bigl[\Delta(\theta)\bigr]\bigr\}\Bigr),$$

where

$$f(\lambda,\overline{r},d) = \bigl[2A(\lambda)\bigr]^{-1}\biggl\{2\overline{r}A(\lambda) - 1 + \sqrt{\bigl[1 - 2\overline{r}A(\lambda)\bigr]^2 + 4A(\lambda)\Bigl\{\overline{r}\bigl[1 - A(\lambda)\bigr] - \tfrac{d}{\lambda}\Bigr\}}\biggr\}.$$

Thus we can find some exchangeable function $\lambda(\omega,\theta)$, such that

$$f\Bigl(\lambda(\omega,\theta), \overline{r}(\theta), -\log\bigl\{\epsilon\pi\bigl[\Delta(\theta)\bigr]\bigr\}\Bigr) = \sup_{\beta\in\mathbb{R}_+} f\Bigl(\beta, \overline{r}(\theta), -\log\bigl\{\epsilon\pi\bigl[\Delta(\theta)\bigr]\bigr\}\Bigr).$$

Applying Corollary 3.2.7 (page 120) to that choice of λ, we see that

THEOREM 3.2.8. *For any exchangeable probability measure $\mathbb{P} \in \mathcal{M}_+^1(\Omega)$, for any exchangeable posterior probability distribution $\pi : \Omega \to \mathcal{M}_+^1(\Theta)$, with $\mathbb{P}$ probability at least $1 - \epsilon$, for any $\theta \in \Theta$, for any $\lambda \in \mathbb{R}_+$,*

$$\overline{r}(\theta) \leq r_1(\theta) + A(\lambda)\bigl[\overline{r}(\theta) - r_1(\theta) r_2(\theta)\bigr] - \frac{\log\bigl\{\epsilon\pi\bigl[\Delta(\theta)\bigr]\bigr\}}{\lambda},$$

where $A(\lambda)$ is defined by equation (3.2, page 119).

Solving the previous inequality in $r_2(\theta)$, we get

COROLLARY 3.2.9. *Under the same assumptions as in the previous theorem, with* $\mathbb{P}$ *probability at least* $1-\epsilon$*, for any* $\theta \in \Theta$*,*

$$r_2(\theta) \leq \inf_{\lambda \in \mathbb{R}_+} \frac{r_1(\theta)\Big\{1 + \frac{2N}{\lambda}\log\big[\cosh(\frac{\lambda}{2N})\big]\Big\} - \frac{2\log\{\epsilon\pi[\Delta(\theta)]\}}{\lambda}}{1 - \frac{2N}{\lambda}\log\big[\cosh(\frac{\lambda}{2N})\big]\big[1 - 2r_1(\theta)\big]}.$$

Applying this to our usual numerical example of a binary classification model with Vapnik–Cervonenkis dimension not greater than $h = 10$, when $N = 1000$, $\inf_\Theta r_1 = r_1(\widehat{\theta}) = 10$ and $\epsilon = 0.01$, we obtain that $r_2(\widehat{\theta}) \leq 0.4450$.

3.3. Vapnik bounds for inductive classification

3.3.1. Arbitrary shadow sample size. We assume in this section that

$$\mathbb{P} = \left(\bigotimes_{i=1}^{N} P_i\right)^{\otimes\infty} \in \mathcal{M}_+^1\Big\{\big[(\mathcal{X}\times\mathcal{Y})^N\big]^{\mathbb{N}}\Big\},$$

where $P_i \in \mathcal{M}_+^1(\mathcal{X}\times\mathcal{Y})$: we consider an infinite i.i.d. sequence of independent *non*-identically distributed samples of size N, the first one only being observed. More precisely, under $\mathbb{P}$ each sample $(X_{i+jN}, Y_{i+jN})_{i=1}^N$ is distributed according to $\bigotimes_{i=1}^N P_i$, and they are all independent from each other. Only the first sample $(X_i, Y_i)_{i=1}^N$ is assumed to be observed. The shadow samples will only appear in the proofs. The aim of this section is to prove better Vapnik bounds, generalizing them in the same time to the independent non-i.i.d. setting, which to our knowledge has not been done before.

Let us introduce the notation $\mathbb{P}'\big[h(\omega)\big] = \mathbb{P}\big[h(\omega)\,|\,(X_i,Y_i)_{i=1}^N\big]$, where h may be any suitable (e.g. bounded) random variable, let us also put $\Omega = \big[(\mathcal{X}\times\mathcal{Y})^N\big]^{\mathbb{N}}$.

DEFINITION 3.3.1. For any subset $A \subset \mathbb{N}$ of integers, let $\mathfrak{C}(A)$ be the set of circular permutations of the totally ordered set A, extended to a permutation of $\mathbb{N}$ by taking it to be the identity on the complement $\mathbb{N}\setminus A$ of A. We will say that a random function $h : \Omega \to \mathbb{R}$ is k-partially exchangeable if

$$h(\omega\circ s) = h(\omega), \quad s \in \mathfrak{C}\big(\{i+jN\,;\,j=0,\dots,k\}\big), i = 1,\dots,N.$$

In the same way, we will say that a posterior distribution $\pi : \Omega \to \mathcal{M}_+^1(\Theta)$ is k-partially exchangeable if

$$\pi(\omega\circ s) = \pi(\omega) \in \mathcal{M}_+^1(\Theta), \quad s \in \mathfrak{C}\big(\{i+jN\,;\,j=0,\dots,k\}\big), i = 1,\dots,N.$$

Note that $\mathbb{P}$ itself is k-partially exchangeable for any k in the sense that for any bounded measurable function $h : \Omega \to \mathbb{R}$

$$\mathbb{P}\big[h(\omega\circ s)\big] = \mathbb{P}\big[h(\omega)\big], \quad s \in \mathfrak{C}\big(\{i+jN\,;\,j=0,\dots,k\}\big), i = 1,\dots,N.$$

Let $\Delta_k(\theta) = \Big\{\theta' \in \Theta\,;\,\big[f_{\theta'}(X_i)\big]_{i=1}^{(k+1)N} = \big[f_\theta(X_i)\big]_{i=1}^{(k+1)N}\Big\}$, $\theta \in \Theta, k \in \mathbb{N}^*$, and let also $\overline{r}_k(\theta) = \dfrac{1}{(k+1)N}\displaystyle\sum_{i=1}^{(k+1)N} \mathbb{1}\big[f_\theta(X_i) \neq Y_i\big]$. Theorem 3.1.2 shows that for any

positive real parameter λ and any k-partially exchangeable posterior distribution $\pi_k : \Omega \to \mathcal{M}_+^1(\Theta)$,

$$\mathbb{P}\biggl\{\exp\biggl[\sup_{\theta\in\Theta}\lambda\bigl[\Phi_{\frac{\lambda}{N}}(\overline{r}_k)-r_1\bigr]+\log\bigl\{\epsilon\pi_k\bigl[\Delta_k(\theta)\bigr]\bigr\}\biggr]\biggr\}\leq\epsilon.$$

Using the general fact that

$$\mathbb{P}\bigl[\exp(h)\bigr]=\mathbb{P}\Bigl\{\mathbb{P}'\bigl[\exp(h)\bigr]\Bigr\}\geq\mathbb{P}\Bigl\{\exp\bigl[\mathbb{P}'(h)\bigr]\Bigr\},$$

and the fact that the expectation of a supremum is larger than the supremum of an expectation, we see that with $\mathbb{P}$ probability at most $1-\epsilon$, for any $\theta\in\Theta$,

$$\mathbb{P}'\Bigl\{\Phi_{\frac{\lambda}{N}}\bigl[\overline{r}_k(\theta)\bigr]\Bigr\}\leq r_1(\theta)-\frac{\mathbb{P}'\Bigl\{\log\bigl\{\epsilon\pi_k\bigl[\Delta_k(\theta)\bigr]\bigr\}\Bigr\}}{\lambda}.$$

For short let us put

$$\begin{aligned}\bar{d}_k(\theta)&=-\log\bigl\{\epsilon\pi_k\bigl[\Delta_k(\theta)\bigr]\bigr\},\\ d'_k(\theta)&=-\mathbb{P}'\Bigl\{\log\bigl\{\epsilon\pi_k\bigl[\Delta_k(\theta)\bigr]\bigr\}\Bigr\},\\ d_k(\theta)&=-\mathbb{P}\Bigl\{\log\bigl\{\epsilon\pi_k\bigl[\Delta_k(\theta)\bigr]\bigr\}\Bigr\}.\end{aligned}$$

We can use the convexity of $\Phi_{\frac{\lambda}{N}}$ and the fact that $\mathbb{P}'(\overline{r}_k)=\frac{r_1+kR}{k+1}$, to establish that

$$\mathbb{P}'\Bigl\{\Phi_{\frac{\lambda}{N}}\bigl[\overline{r}_k(\theta)\bigr]\Bigr\}\geq\Phi_{\frac{\lambda}{N}}\left[\frac{r_1(\theta)+kR(\theta)}{k+1}\right].$$

We have proved

THEOREM 3.3.1. *Using the above hypotheses and notation, for any sequence $\pi_k : \Omega\to\mathcal{M}_+^1(\Theta)$, where π_k is a k-partially exchangeable posterior distribution, for any positive real constant λ, any positive integer k, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta\in\Theta$,*

$$\Phi_{\frac{\lambda}{N}}\left[\frac{r_1(\theta)+kR(\theta)}{k+1}\right]\leq r_1(\theta)+\frac{d'_k(\theta)}{\lambda}.$$

We can make as we did with Theorem 1.2.6 (page 11) the result of this theorem uniform in $\lambda\in\{\alpha^j\,;\,j\in\mathbb{N}^*\}$ and $k\in\mathbb{N}^*$ (considering on k the prior $\frac{1}{k(k+1)}$ and on j the prior $\frac{1}{j(j+1)}$), and obtain

THEOREM 3.3.2. *For any real parameter $\alpha>1$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta\in\Theta$,*

$$R(\theta)\leq\inf_{k\in\mathbb{N}^*,j\in\mathbb{N}^*}\frac{1-\exp\Bigl\{-\frac{\alpha^j}{N}r_1(\theta)-\frac{1}{N}\Bigl\{d'_k(\theta)+\log\bigl[k(k+1)j(j+1)\bigr]\Bigr\}\Bigr\}}{\frac{k}{k+1}\left[1-\exp\left(-\frac{\alpha^j}{N}\right)\right]}-\frac{r_1(\theta)}{k}.$$

As a special case we can choose π_k such that $\log\{\pi_k[\Delta_k(\theta)]\}$ is independent of θ and equal to $\log(\mathfrak{N}_k)$, where

$$\mathfrak{N}_k = \left|\left\{ \left[f_\theta(X_i)\right]_{i=1}^{(k+1)N} ; \theta \in \Theta\right\}\right|$$

is the size of the trace of the classification model on the extended sample of size $(k+1)N$. With this choice, we obtain a bound involving a new flavour of conditional Vapnik entropy, namely

$$d'_k(\theta) = \mathbb{P}\left[\log(\mathfrak{N}_k)\,|(Z_i)_{i=1}^N\right] - \log(\epsilon).$$

In the case of binary classification using a Vapnik–Cervonenkis class of Vapnik–Cervonenkis dimension not greater than $h = 10$, when $N = 1000$, $\inf_\Theta r_1 = r_1(\widehat{\theta}) = 0.2$ and $\epsilon = 0.01$, choosing $\alpha = 1.1$, we obtain $R(\widehat{\theta}) \leq 0.4271$ (for an optimal value of $\lambda = 1071.8$, and an optimal value of $k = 16$).

3.3.2. A better minimization with respect to the exponential parameter. If we are not pleased with optimizing λ on a discrete subset of the real line, we can use a slightly different approach. From Theorem 3.1.2 (page 113), we see that for any positive integer k, for any k-partially exchangeable positive real measurable function $\lambda : \Omega \times \Theta \to \mathbb{R}_+$ satisfying equation (3.1, page 116) — with $\Delta(\theta)$ replaced with $\Delta_k(\theta)$ — for any $\epsilon \in]0,1)$ and $\eta \in]0,1)$,

$$\mathbb{P}\left\{\mathbb{P}'\left[\exp\left[\sup_\theta \lambda\left[\Phi_{\frac{\lambda}{N}}(\overline{r}_k) - r_1\right] + \log\left\{\epsilon\eta\pi_k\left[\Delta_k(\theta)\right]\right\}\right]\right\} \leq \epsilon\eta,$$

therefore with $\mathbb{P}$ probability at least $1-\epsilon$,

$$\mathbb{P}'\left\{\exp\left[\sup_\theta \lambda\left[\Phi_{\frac{\lambda}{N}}(\overline{r}_k) - r_1\right] + \log\left\{\epsilon\eta\pi_k\left[\Delta_k(\theta)\right]\right\}\right]\right\} \leq \eta,$$

and consequently, with $\mathbb{P}$ probability at least $1-\epsilon$, with $\mathbb{P}'$ probability at least $1-\eta$, for any $\theta \in \Theta$,

$$\Phi_{\frac{\lambda}{N}}(\overline{r}_k) + \frac{\log\left\{\epsilon\eta\pi_k\left[\Delta_k(\theta)\right]\right\}}{\lambda} \leq r_1.$$

Now we are entitled to choose

$$\lambda(\omega,\theta) \in \arg\max_{\lambda' \in \mathbb{R}_+} \Phi_{\frac{\lambda'}{N}}(\overline{r}_k) + \frac{\log\left\{\epsilon\eta\pi_k\left[\Delta_k(\theta)\right]\right\}}{\lambda'}.$$

This shows that with $\mathbb{P}$ probability at least $1-\epsilon$, with $\mathbb{P}'$ probability at least $1-\eta$, for any $\theta \in \Theta$,

$$\sup_{\lambda \in \mathbb{R}_+} \Phi_{\frac{\lambda}{N}}(\overline{r}_k) - \frac{\bar{d}_k(\theta) - \log(\eta)}{\lambda} \leq r_1,$$

which can also be written

$$\Phi_{\frac{\lambda}{N}}(\overline{r}_k) - r_1 - \frac{\bar{d}_k(\theta)}{\lambda} \leq -\frac{\log(\eta)}{\lambda}, \quad \lambda \in \mathbb{R}_+.$$

Thus with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$, any $\lambda \in \mathbb{R}_+$,

$$\mathbb{P}'\left[\Phi_{\frac{\lambda}{N}}(\overline{r}_k) - r_1 - \frac{\bar{d}_k(\theta)}{\lambda}\right] \leq -\frac{\log(\eta)}{\lambda} + \left[1 - r_1 + \frac{\log(\eta)}{\lambda}\right]\eta.$$

On the other hand, $\Phi_{\frac{\lambda}{N}}$ being a convex function,

$$\mathbb{P}'\left[\Phi_{\frac{\lambda}{N}}(\overline{r}_k) - r_1 - \frac{\bar{d}_k(\theta)}{\lambda}\right] \geq \Phi_{\frac{\lambda}{N}}\left[\mathbb{P}'(\overline{r}_k)\right] - r_1 - \frac{d'_k}{\lambda}$$
$$= \Phi_{\frac{\lambda}{N}}\left(\frac{kR + r_1}{k+1}\right) - r_1 - \frac{d'_k}{\lambda}.$$

Thus with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$,

$$\frac{kR + r_1}{k+1} \leq \inf_{\lambda \in \mathbb{R}_+} \Phi^{-1}_{\frac{\lambda}{N}}\left[r_1(1-\eta) + \eta + \frac{d'_k - \log(\eta)(1-\eta)}{\lambda}\right].$$

We can generalize this approach by considering a finite decreasing sequence $\eta_0 = 1 > \eta_1 > \eta_2 > \cdots > \eta_J > \eta_{J+1} = 0$, and the corresponding sequence of levels

$$L_j = -\frac{\log(\eta_j)}{\lambda}, 0 \leq j \leq J,$$
$$L_{J+1} = 1 - r_1 - \frac{\log(J) - \log(\epsilon)}{\lambda}.$$

Taking a union bound in j, we see that with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$, for any $\lambda \in \mathbb{R}_+$,

$$\mathbb{P}'\left[\Phi_{\frac{\lambda}{N}}(\overline{r}_k) - r_1 - \frac{\bar{d}_k + \log(J)}{\lambda} \geq L_j\right] \leq \eta_j, \quad j = 0, \ldots, J+1,$$

and consequently

$$\mathbb{P}'\left[\Phi_{\frac{\lambda}{N}}(\overline{r}_k) - r_1 - \frac{\bar{d}_k + \log(J)}{\lambda}\right]$$
$$\leq \int_0^{L_{J+1}} \mathbb{P}'\left[\Phi_{\frac{\lambda}{N}}(\overline{r}_k) - r_1 - \frac{\bar{d}_k + \log(J)}{\lambda} \geq \alpha\right] d\alpha \quad \leq \sum_{j=1}^{J+1} \eta_{j-1}(L_j - L_{j-1})$$
$$= \eta_J\left[1 - r_1 - \frac{\log(J) - \log(\epsilon) - \log(\eta_J)}{\lambda}\right] - \frac{\log(\eta_1)}{\lambda} + \sum_{j=1}^{J-1} \frac{\eta_j}{\lambda} \log\left(\frac{\eta_j}{\eta_{j+1}}\right).$$

Let us put

$$d''_k\left[\theta, (\eta_j)_{j=1}^J\right] = d'_k(\theta) + \log(J) - \log(\eta_1)$$
$$+ \sum_{j=1}^{J-1} \eta_j \log\left(\frac{\eta_j}{\eta_{j+1}}\right) + \log\left(\frac{\epsilon \eta_J}{J}\right)\eta_J.$$

We have proved that for any decreasing sequence $(\eta_j)_{j=1}^J$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$,

$$\frac{kR + r_1}{k+1} \leq \inf_{\lambda \in \mathbb{R}_+} \Phi^{-1}_{\frac{\lambda}{N}}\left[r_1(1-\eta_J) + \eta_J + \frac{d''_k\left[\theta, (\eta_j)_{j=1}^J\right]}{\lambda}\right].$$

Remark 3.3.1. We can for instance choose $J = 2$, $\eta_2 = \frac{1}{10N}$, $\eta_1 = \frac{1}{\log(10N)}$, resulting in

$$d''_k = d'_k + \log(2) + \log\log(10N) + 1 - \frac{\log\log(10N)}{\log(10N)} - \frac{\log\left(\frac{20N}{\epsilon}\right)}{10N}.$$

In the case where $N = 1000$ and for any $\epsilon \in]0,1)$, we get $d''_k \leq d'_k + 3.7$, in the case where $N = 10^6$, we get $d''_k \leq d'_k + 4.4$, and in the case $N = 10^9$, we get $d''_k \leq d'_k + 4.7$.

Therefore, for any practical purpose we could take $d''_k = d'_k + 4.7$ and $\eta_J = \frac{1}{10N}$ in the above inequality.

Taking moreover a weighted union bound in k, we get

THEOREM 3.3.3. *For any $\epsilon \in]0,1)$, any sequence $1 > \eta_1 > \cdots > \eta_J > 0$, any sequence $\pi_k : \Omega \to \mathcal{M}_+^1(\Theta)$, where π_k is a k-partially exchangeable posterior distribution, with $\mathbb{P}$ probability at least $1 - \epsilon$, for any $\theta \in \Theta$,*

$$R(\theta) \leq \inf_{k \in \mathbb{N}^*} \frac{k+1}{k} \inf_{\lambda \in \mathbb{R}_+} \Phi_{\frac{\lambda}{N}}^{-1} \Bigg[r_1(\theta) + \eta_J \big[1 - r_1(\theta)\big] + \frac{d''_k\big[\theta, (\eta_j)_{j=1}^J\big] + \log\big[k(k+1)\big]}{\lambda} \Bigg] - \frac{r_1(\theta)}{k}.$$

COROLLARY 3.3.4. *For any $\epsilon \in]0,1)$, for any $N \leq 10^9$, with $\mathbb{P}$ probability at least $1 - \epsilon$, for any $\theta \in \Theta$,*

$$R(\theta) \leq \inf_{k \in \mathbb{N}^*} \inf_{\lambda \in \mathbb{R}_+} \frac{k+1}{k} \big[1 - \exp(-\tfrac{\lambda}{N})\big]^{-1} \Bigg\{ 1 - \exp\Bigg[-\tfrac{\lambda}{N} \Big[r_1(\theta) + \tfrac{1}{10N} \Big] - \frac{\mathbb{P}'\big[\log(\mathfrak{N}_k) \,|\, (Z_i)_{i=1}^N\big] - \log(\epsilon) + \log\big[k(k+1)\big] + 4.7}{N} \Bigg] \Bigg\} - \frac{r_1(\theta)}{k}.$$

Let us end this section with a numerical example: in the case of binary classification with a Vapnik–Cervonenkis class of dimension not greater than 10, when $N = 1000$, $\inf_\Theta r_1 = r_1(\widehat{\theta}) = 0.2$ and $\epsilon = 0.01$, we get a bound $R(\widehat{\theta}) \leq 0.4211$ (for optimal values of $k = 15$ and of $\lambda = 1010$).

3.3.3. EQUAL SHADOW AND TRAINING SAMPLE SIZES. In the case when $k = 1$, we can use Theorem 3.2.5 (page 118) and replace $\Phi_{\frac{\lambda}{N}}^{-1}(q)$ with $\big\{1 - \frac{2N}{\lambda} \times \log\big[\cosh(\frac{\lambda}{2N})\big]\big\}^{-1} q$, resulting in

THEOREM 3.3.5. *For any $\epsilon \in]0,1)$, any $N \leq 10^9$, any one-partially exchangeable posterior distribution $\pi_1 : \Omega \to \mathcal{M}_+^1(\Theta)$, with $\mathbb{P}$ probability at least $1 - \epsilon$, for any $\theta \in \Theta$,*

$$R(\theta) \leq \inf_{\lambda \in \mathbb{R}_+} \frac{\Big\{1 + \frac{2N}{\lambda} \log\big[\cosh(\frac{\lambda}{2N})\big]\Big\} r_1(\theta) + \frac{1}{5N} + 2\frac{d'_1(\theta) + 4.7}{\lambda}}{1 - \frac{2N}{\lambda} \log\big[\cosh(\frac{\lambda}{2N})\big]}.$$

3.3.4. IMPROVEMENT ON THE EQUAL SAMPLE SIZE BOUND IN THE I.I.D. CASE. Finally, in the case when $\mathbb{P}$ is i.i.d., meaning that all the P_i are equal, we can improve the previous bound. For any partially exchangeable function $\lambda : \Omega \times \Theta \to \mathbb{R}_+$, we saw in the discussion preceding Theorem 3.2.6 (page 120) that

$$T\Big[\exp\big[\lambda(\overline{r}_k - r_1) - A(\lambda)v\big]\Big] \leq 1,$$

with the notation introduced therein. Thus for any partially exchangeable positive real measurable function $\lambda : \Omega \times \Theta \to \mathbb{R}_+$ satisfying equation (3.1, page 116), any one-partially exchangeable posterior distribution $\pi_1 : \Omega \to \mathcal{M}_+^1(\Theta)$,

$$\mathbb{P}\Big\{\exp\Big[\sup_{\theta\in\Theta} \lambda\big[\overline{r}_k(\theta) - r_1(\theta) - A(\lambda)v(\theta)\big] + \log\big[\epsilon\pi_1\big[\Delta(\theta)\big]\Big]\Big\} \leq 1.$$

Therefore with $\mathbb{P}$ probability at least $1-\epsilon$, with $\mathbb{P}'$ probability $1-\eta$,

$$\overline{r}_k(\theta) \leq r_1(\theta) + A(\lambda)v(\theta) + \frac{1}{\lambda}\big[\bar{d}_1(\theta) - \log(\eta)\big].$$

We can then choose $\lambda(\omega,\theta) \in \arg\min_{\lambda'\in\mathbb{R}_+} A(\lambda')v(\theta) + \dfrac{\bar{d}_1(\theta) - \log(\eta)\big]}{\lambda'}$, which satisfies the required conditions, to show that with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$, with $\mathbb{P}'$ probability at least $1-\eta$, for any $\lambda \in \mathbb{R}_+$,

$$\overline{r}_k(\theta) \leq r_1(\theta) + A(\lambda)v(\theta) + \frac{\bar{d}_1(\theta) - \log(\eta)}{\lambda}.$$

We can then take a union bound on a decreasing sequence of J values $\eta_1 \geq \cdots \geq \eta_J$ of η. Weakening the order of quantifiers a little, we then obtain the following statement: with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$, for any $\lambda \in \mathbb{R}_+$, for any $j = 1, \ldots, J$

$$\mathbb{P}'\left[\overline{r}_k(\theta) - r_1(\theta) - A(\lambda)v(\theta) - \frac{\bar{d}_1(\theta) + \log(J)}{\lambda} \geq -\frac{\log(\eta_j)}{\lambda}\right] \leq \eta_j.$$

Consequently for any $\lambda \in \mathbb{R}_+$,

$$\begin{aligned}\mathbb{P}'\left[\overline{r}_k(\theta) - r_1(\theta) - A(\lambda)v(\theta) - \frac{\bar{d}_1(\theta) + \log(J)}{\lambda}\right] \\ \leq -\frac{\log(\eta_1)}{\lambda} + \eta_J\left[1 - r_1(\theta) - \frac{\log(J) - \log(\epsilon) - \log(\eta_J)}{\lambda}\right] \\ + \sum_{j=1}^{J-1} \frac{\eta_j}{\lambda}\log\left(\frac{\eta_j}{\eta_{j+1}}\right).\end{aligned}$$

Moreover $\mathbb{P}'\big[v(\theta)\big] = \frac{r_1+R}{2} - r_1 R$, (this is where we need equidistribution) thus proving that

$$\frac{R - r_1}{2} \leq \frac{A(\lambda)}{2}\Big[R + r_1 - 2r_1 R\Big] + \frac{d_1''\big[\theta, (\eta_j)_{j=1}^J\big]}{\lambda} + \eta_J\big[1 - r_1(\theta)\big].$$

Keeping track of quantifiers, we obtain

THEOREM 3.3.6. *For any decreasing sequence $(\eta_j)_{j=1}^J$, any $\epsilon \in]0,1)$, any one-partially exchangeable posterior distribution $\pi : \Omega \to \mathcal{M}_+^1(\Theta)$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$,*

$$R(\theta) \leq \inf_{\lambda\in\mathbb{R}_+} \frac{\Big\{1 + \frac{2N}{\lambda}\log\big[\cosh(\frac{\lambda}{2N})\big]\Big\}r_1(\theta) + \dfrac{2d_1''\big[\theta,(\eta_j)_{j=1}^J\big]}{\lambda} + 2\eta_J\big[1 - r_1(\theta)\big]}{1 - \frac{2N}{\lambda}\log\big[\cosh(\frac{\lambda}{2N})\big]\big[1 - 2r_1(\theta)\big]}.$$

3.4. Gaussian approximation in Vapnik bounds

3.4.1. Gaussian upper bounds of variance terms. To obtain formulas which could be easily compared with original Vapnik bounds, we may replace $p - \Phi_a(p)$ with a Gaussian upper bound:

Lemma 3.4.1. *For any $p \in (0, \frac{1}{2})$, any $a \in \mathbb{R}_+$,*

$$p - \Phi_a(p) \leq \frac{a}{2} p(1-p).$$

For any $p \in (\frac{1}{2}, 1)$,

$$p - \Phi_a(p) \leq \frac{a}{8}.$$

Proof. Let us notice that for any $p \in (0,1)$,

$$\begin{aligned}
\frac{\partial}{\partial a}\big[-a\Phi_a(p)\big] &= -\frac{p\exp(-a)}{1-p+p\exp(-a)}, \\
\frac{\partial^2}{\partial^2 a}\big[-a\Phi_a(p)\big] &= \frac{p\exp(-a)}{1-p+p\exp(-a)}\left(1 - \frac{p\exp(-a)}{1-p+p\exp(-a)}\right) \\
&\leq \begin{cases} p(1-p) & p \in (0, \frac{1}{2}), \\ \frac{1}{4} & p \in (\frac{1}{2}, 1). \end{cases}
\end{aligned}$$

Thus taking a Taylor expansion of order one with integral remainder:

$$-a\Phi(a) \leq \begin{cases} -ap + \displaystyle\int_0^a p(1-p)(a-b)db & \\ \qquad = -ap + \dfrac{a^2}{2} p(1-p), & p \in (0, \frac{1}{2}), \\ -ap + \displaystyle\int_0^a \frac{1}{4}(a-b)db = -ap + \dfrac{a^2}{8}, & p \in (\frac{1}{2}, 1). \end{cases}$$

This ends the proof of our lemma. □

Lemma 3.4.2. *Let us consider the bound*

$$B(q,d) = \left(1 + \frac{2d}{N}\right)^{-1} \left[q + \frac{d}{N} + \sqrt{\frac{2dq(1-q)}{N} + \frac{d^2}{N^2}}\right], \quad q \in \mathbb{R}_+, d \in \mathbb{R}_+.$$

Let us also put

$$\bar{B}(q,d) = \begin{cases} B(q,d) & B(q,d) \leq \frac{1}{2}, \\ q + \sqrt{\frac{d}{2N}} & \text{otherwise.} \end{cases}$$

For any positive real parameters q and d

$$\inf_{\lambda \in \mathbb{R}_+} \Phi^{-1}_{\frac{\lambda}{N}}\left(q + \frac{d}{\lambda}\right) \leq \bar{B}(q,d).$$

Proof. Let $p = \inf\limits_{\lambda} \Phi^{-1}_{\frac{\lambda}{N}}\left(q + \dfrac{d}{\lambda}\right)$. For any $\lambda \in \mathbb{R}_+$,

$$p - \frac{\lambda}{2N}(p \wedge \tfrac{1}{2})\big[1 - (p \wedge \tfrac{1}{2})\big] \leq \Phi_{\frac{\lambda}{N}}(p) \leq q + \frac{d}{\lambda}.$$

Thus

$$p \leq q + \inf_{\lambda \in \mathbb{R}_+} \frac{\lambda}{2N}(p \wedge \tfrac{1}{2})\big[1 - (p \wedge \tfrac{1}{2})\big] + \frac{d}{\lambda}$$
$$= q + \sqrt{\frac{2d(p \wedge \frac{1}{2})\big[1 - (p \wedge \frac{1}{2})\big]}{N}} \leq q + \sqrt{\frac{d}{2N}}.$$

Then let us remark that $B(q,d) = \sup\left\{p' \in \mathbb{R}_+ \,;\, p' \leq q + \sqrt{\frac{2dp'(1-p')}{N}}\right\}$. If moreover $\frac{1}{2} \geq B(q,d)$, then according to this remark $\frac{1}{2} \geq q + \sqrt{\frac{d}{2N}} \geq p$. Therefore $p \leq \frac{1}{2}$, and consequently $p \leq q + \sqrt{\frac{2dp(1-p)}{N}}$, implying that $p \leq B(q,d)$. □

3.4.2. Arbitrary shadow sample size. The previous lemma combined with Corollary 3.3.4 (page 125) implies

Corollary 3.4.3. *Let us use the notation introduced in Lemma 3.4.2 (page 127). For any $\epsilon \in]0,1)$, any integer $N \leq 10^9$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$,*

$$R(\theta) \leq \inf_{k \in \mathbb{N}^*} \frac{k+1}{k}\left\{\bar{B}\Big[r_1(\theta) + \frac{1}{10N}, d'_k(\theta) + \log\big[k(k+1)\big] + 4.7\Big]\right\} - \frac{r_1(\theta)}{k}.$$

3.4.3. Equal sample sizes in the i.i.d. case. To make a link with Vapnik's result, it is useful to state the Gaussian approximation to Theorem 3.3.6 (page 126). Indeed, using the upper bound $A(\lambda) \leq \frac{\lambda}{4N}$, where $A(\lambda)$ is defined by equation (3.2) on page 119, we get with $\mathbb{P}$ probability at least $1-\epsilon$

$$R - r_1 - 2\eta_J \leq \inf_{\lambda \in \mathbb{R}_+} \frac{\lambda}{4N}\big[R + r_1 - 2r_1 R\big] + \frac{2d''_1}{\lambda} = \sqrt{\frac{2d''_1(R + r_1 - 2r_1 R)}{N}},$$

which can be solved in R to obtain

Corollary 3.4.4. *With $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$,*

$$R(\theta) \leq r_1(\theta) + \frac{d''_1(\theta)}{N}\big[1 - 2r_1(\theta)\big] + 2\eta_J$$
$$+ \sqrt{\frac{4d''_1(\theta)\big[1 - r_1(\theta)\big]r_1(\theta)}{N} + \frac{d''_1(\theta)^2}{N^2}\big[1 - 2r_1(\theta)\big]^2 + \frac{4d''_1(\theta)}{N}\big[1 - 2r_1(\theta)\big]\eta_J}.$$

This is to be compared with Vapnik's result, as proved in Vapnik (1998, page 138):

Theorem 3.4.5 (Vapnik). *For any i.i.d. probability distribution $\mathbb{P}$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$, putting*

$$d_V = \log\big[\mathbb{P}(\mathfrak{N}_1)\big] + \log(4/\epsilon),$$

$$R(\theta) \leq r_1(\theta) + \frac{2d_V}{N} + \sqrt{\frac{4d_V r_1(\theta)}{N} + \frac{4d_V^2}{N^2}}.$$

Recalling that we can choose $(\eta_j)_{j=1}^2$ such that $\eta_J = \eta_2 = \frac{1}{10N}$ (which brings a negligible contribution to the bound) and such that for any $N \leq 10^9$,

$$d_1''(\theta) \leq \mathbb{P}\big[\log(\mathfrak{N}_1) \,|\, (Z_i)_{i=1}^N\big] - \log(\epsilon) + 4.7,$$

we see that our complexity term is somehow more satisfactory than Vapnik's, since it is integrated outside the logarithm, with a slightly larger additional constant (remember that $\log 4 \simeq 1.4$, which is better than our 4.7, which could presumably be improved by working out a better sequence η_j, but not down to $\log(4)$). Our variance term is better, since we get $r_1(1-r_1)$, instead of r_1. We also have $\dfrac{d_1''}{N}$ instead of $2\dfrac{d_V}{N}$, because we use no symmetrization trick.

Let us illustrate these bounds on a numerical example, corresponding to a situation where the sample is noisy or the classification model is weak. Let us assume that $N = 1000$, $\inf_\Theta r_1 = r_1(\widehat{\theta}) = 0.2$, that we are performing binary classification with a model with Vapnik–Cervonenkis dimension not greater than $h = 10$, and that we work at confidence level $\epsilon = 0.01$. Vapnik's theorem provides an upper bound for $R(\widehat{\theta})$ not smaller than 0.610, whereas Corollary 3.4.4 gives $R(\widehat{\theta}) \leq 0.461$ (using the bound $d_1'' \leq d_1' + 3.7$ when $N = 1000$). Now if we go for Theorem 3.3.6 and do not make a Gaussian approximation, we get $R(\widehat{\theta}) \leq 0.453$. It is interesting to remark that this bound is achieved for $\lambda = 1195 > N = 1000$. This explains why the Gaussian approximation in Vapnik's bound can be improved: for such a large value of λ, $\lambda r_1(\theta)$ does not behave like a Gaussian random variable.

Let us recall in conclusion that the best bound is provided by Theorem 3.3.3 (page 125), giving $R(\widehat{\theta}) \leq 0.4211$, (that is approximately 2/3 of Vapnik's bound), for optimal values of $k = 15$, and of $\lambda = 1010$. This bound can be seen to take advantage of the fact that Bernoulli random variables are not Gaussian (its Gaussian approximation, Corollary 3.4.3, gives a bound $R(\theta) \simeq 0.4325$, still with an optimal $k = 15$), and of the fact that the optimal size of the shadow sample is significantly larger than the size of the observed sample. Moreover, Theorem 3.3.3 does not assume that the sample is i.i.d., but only that it is independent, thus generalizing Vapnik's bounds to inhomogeneous data (this will presumably be the case when data are collected from different places where the experimental conditions may not be the same, although they may reasonably be assumed to be independent).

Our little numerical example was chosen to illustrate the case when it is nontrivial to decide whether the chosen classifier does better than the 0.5 error rate of blind random classification. This case is of interest to choose "weak learners" to be aggregated or combined in some appropriate way in a second stage to reach a better classification rate. This stage of feature selection is unavoidable in many real world classification tasks. Our little computations are meant to exemplify the fact that Vapnik's bounds, although asymptotically suboptimal, as is obvious by comparison with the first two chapters, can do the job when dealing with moderate sample sizes.