Transductive PAC-Bayesian classification

Olivier Catoni – CNRS Université Paris 6
July 20 2004
– Gatsby Computational Neuroscience Unit–
University College London

catoni@ccr.jussieu.fr
Transductive PAC-Bayesian theorems, an introduction

- \((\mathcal{X}, \mathcal{B})\) a measurable set of patterns to be classified;
- \(\mathcal{Y}\) a finite set of labels, applied to the patterns (most of the time, we will consider the binary case \(\mathcal{Y} = \{0, 1\}\));
- \((X_i, Y_i)_{i=1}^{N+M} \overset{\text{def}}{=} (Z_i)_{i=1}^{N+M} \overset{\text{notation}}{=} Z_1^{N+M}\), the canonical process on \((\mathcal{X} \times \mathcal{Y})^{N+M}\);
- \(\mathcal{R} = \{f_\theta : \mathcal{X} \rightarrow \mathcal{Y} : \theta \in \Theta\}\) some family (or union of families) of classification rules;
- \(\mathbf{P}\) some joint distribution on \((\mathcal{X} \times \mathcal{Y})^{N+M}\);
"Classical" PAC bounds: $M = 0$, $\mathbf{P}$ is a product measure: $\mathbf{P} = P^{\otimes N}$, and $R(\theta) = P[f_{\theta}(X) \neq Y]$ is to be compared with

$$r(\theta) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[f_{\theta}(X_i) \neq Y_i],$$

through an inequality of the type:

With $\mathbf{P} = P^{\otimes N}$ probability at least $1 - \epsilon$, for any $\theta \in \Theta$,

$$R(\theta) \leq r(\theta) + \gamma(\theta),$$

where $\gamma(\theta)$ depends only on $Z_1^N$ and not directly on $\mathbf{P}$.

Intended use of the bound:
- build an estimator by minimizing $r(\theta) + \gamma(\theta)$ in $\theta$;
- more generally, bound the generalization error of any given estimator at some level of confidence $\epsilon$. 
Extensions of this classical setting:

- Putting things into a pseudo Bayesian perspective: replace $R(\theta)$ with $\rho(R)$, where $\rho \in \mathcal{M}_1^+(\Theta)$ ranges into the posterior probability measures on the parameter space $\Theta$ ($\rho$ is allowed to depend on $Z_i^N$). Look for a PAC Bayesian bound of the form: With $\mathbb{P}$ probability at least $1 - \epsilon$, for any $\rho \in \mathcal{M}_1^+(\Theta)$,

$$\rho(R) \leq \rho(r) + \gamma(\rho).$$

There is no universal choice of $\gamma(\rho)$, and one way to choose one penalty function $\gamma$ is to relate $\gamma(\rho)$ with a prior distribution $\pi \in \mathcal{M}_1^+$, independent of $\mathbb{P}$ and of $Z_i^N$. One advantage of the pseudo Bayesian setting is that we can get explicit penalties $\gamma_\pi(\rho)$, where the “complexity” of the model is captured through $\mathcal{K}(\rho, \pi)$. 
Another one is that we can always take $\rho$ to be a finite convex combination of $\mathbb{1}(\theta \in \Lambda))\pi(\Lambda)^{-1}\pi$, where $\Lambda$ ranges into the components of $\Theta$ under the relation

$\theta \sim \theta' \iff f_{\theta}(X_i) = f_{\theta'}(X_i), i = 1, \ldots, N$, or even with the coarser relation $\theta \sim \theta' \iff r(\theta) = r(\theta')$. Doing this, we show that the parameter space can always be reduced to a finite dimensional one, with maximum dimension $2^N$ (in the binary case), although this reduction is data dependent: this is a first step towards Vapnik’s point of view.
- The transductive point of view: $M > 0$, introducing a test set $(X_{N+1}, \ldots, X_{N+M})$. Use the new notation $r_1(\theta)$ for $r(\theta)$, and introduce

$$r_2(\theta) = \frac{1}{M} \sum_{i=N+1}^{N+M} 1[f_\theta(X_i) \neq Y_i].$$

We recover the inductive setting as $M \to +\infty$, since

$$\lim_{M \to +\infty} r_2(\theta) = R(\theta).$$

An interesting case though is when $M = N$.

Interesting features of this approach are:

- Deviation bounds for $r_2(\theta) - r_1(\theta)$ can be obtained under the weaker assumption that $P$ is exchangeable.
Let us put for any $z \in (\mathcal{X} \times \mathcal{Y})^{N+M}$

$$
P_z = \frac{1}{|\mathcal{S}|} \sum_{\sigma} \delta_{z \circ \sigma}.
$$

Any exchangeable distribution $P$ can be decomposed into

$$
P = \int P_z P(dz),
$$

therefore it is enough to prove PAC bounds for $P_z$. 


with the advantage that under $\mathbb{P}_z$:

- the pattern space $\mathcal{X}$ is $\mathbb{P}_z$ almost surely finite (and therefore we have to choose among at most $2^{N+M}$ possible classification rules);

- any exchangeable function is almost surely constant: this allows to consider data dependent priors $\pi$, as long as the dependence on the data is invariant under permutations. This leads to some PAC-Bayesian version of Vapnik’s theory.

- Inductive bounds can be recovered by integrating with respect to the test set.
Transductive PAC Bayesian lemma

Let us consider some regular conditional probability measure \( \pi : X^{N+M} \rightarrow \mathcal{M}_1^1(\Theta) \) and assume that it is exchangeable (i.e. invariant under the permutations of the indices).

The PAC-Bayesian approach starts with an exponential inequality for any fixed value of \( \theta \). We will take \( M = kN \) for convenience.

**Lemma.** For any exchangeable \( \eta : (X \times Y)^{(k+1)N} \times \Theta \rightarrow \mathbb{R} \), for any \( \theta \in \Theta \),

\[
P \left\{ \exp \left[ \lambda \left[ r_2(\theta) - r_1(\theta) \right] - \eta(\theta) \right] \right\} \\
\leq P_{(k+1)N} \left\{ \exp \left[ \frac{\lambda^2}{2N} \left( \frac{1}{k} r_1(\theta) + r_2(\theta) \right) - \eta(\theta) \right] \right\},
\]

(Requires only the invariance under the permutations of \( (i + jN)_{j=0}^k \).)
Let us integrate this inequality with respect to $\pi$ and use the following formula related to the Legendre transform of the Kullback divergence function:

**Lemma.** For any upper bounded measurable function $h$, any probability measure $\rho \in \mathcal{M}^+_1(\Theta, \mathcal{T})$,

$$\log \left\{ \pi \left[ \exp \left( h(\theta) \right) \right] \right\} + K(\rho, \pi) - \rho \left[ h(\theta) \right] = K(\rho, \pi_{\exp(h)}),$$

where $d\pi_{\exp(h)} = \frac{\exp(h)}{\pi[\exp(h)]} d\pi$. 


We obtain the following learning lemma:

**Lemma.** For any exchangeable random variable \( \lambda \in \mathbb{R}_+ \) and any exchangeable threshold \( \eta(\theta) \),

\[
P_{(k+1)N} \left\{ \sup_{\rho \in \mathcal{N}_1(\theta)} \lambda \rho[r_2(\theta)] - \lambda \rho[r_1(\theta)] - \rho[\eta(\theta)] - \mathcal{K}(\rho, \theta) \geq 0 \right\} \\
\leq P_{(k+1)N} \left\{ \pi \left[ \exp \left( \frac{\lambda^2}{2N} \left[ \frac{1}{k} r_1(\theta) + r_2(\theta) \right] - \eta(\theta) \right) \right] \right\}.
\]
We deduce a non localized PAC Bayesian bound by considering
\[ \eta(\theta) = \frac{\lambda^2}{2N} [\frac{k}{N} r_1(\theta) + r_2(\theta)] + \log(\epsilon^{-1}) : \]

**Theorem.** With \( P \) probability at least \( 1 - \epsilon \), for any posterior \( \rho \in \mathcal{M}_1(\Theta) \),

\[
\rho[r_2(\theta)] \leq \left( 1 - \frac{\lambda}{2N} \right)^{-1} \left\{ \left[ 1 + \frac{\lambda}{2kN} \right] \rho[r_1(\theta)] + \frac{\mathcal{X}(\rho, \pi) + \log(\epsilon^{-1})}{\lambda} \right\}.
\]

Considering \( N(X_1^{(k+1)N}) = |\{ f_\theta(X_k) \}_{k=1}^{(k+1)N} : \theta \in \Theta \} | \), the number of traces of \( \{ f_\theta \} \) on \( X_1^{(k+1)N} \), choosing for \( \pi \) the uniform distribution on these traces, and putting

\[
\lambda = \left( \frac{2N \left[ \log[N(X_1^{(k+1)N})] + \log(\epsilon^{-1}) \right]}{k^{-1} r_1(\theta) + r_2(\theta)} \right)^{1/2},
\]
we get

**Corollary.** With probability at least $1 - \epsilon$, for any $\theta \in \Theta$,

$$r_2(\theta) \leq r_1(\theta) + \frac{d}{N} + \sqrt{\frac{2d(1 + k^{-1})r_1(\theta)}{N}} + \frac{d^2}{N^2},$$

where $d = \log \left[ N \left( X_1^{(k+1)N} \right) \right] + \log(\epsilon^{-1})$.

When $Y = \{0, 1\}$,

$$\log \left[ N \left( X_1^{(k+1)N} \right) \right] \leq (k + 1) NH\left( \frac{h}{(k+1)N} \right) \leq h \log\left( \frac{e^{(k+1)N}}{h} \right),$$

where $H(p) = -p \log(p) - (1 - p) \log(1 - p)$ and

$$h = \max \{|A| : A \subset \{X_1^{(k+1)N}\} \text{ and } |A \cap f_\theta^{-1}(1) : \theta \in \Theta| = 2|A|\}$$

is the Vapnik Cervonenkis dimension of the family of classification rules $\{f_\theta : \theta \in \Theta\}$ on the set $\{X_1, \ldots, X_{(k+1)N}\}$. 
In the i.i.d. case when $P = P^{(k+1)N}$, integrating with respect to the test set, we get the following inductive theorem

**Theorem.** With $P^N$ probability at least $1 - \epsilon$, for any $\theta \in \Theta$,

$$R(\theta) \leq r_1(\theta) + \frac{(1 + k^{-1})d^*}{N}$$

$$+ \sqrt{\left( \frac{(1 + k^{-1})d^*}{N} \right)^2 + \frac{2(1 + k^{-1})d^*r_1(\theta)}{N}},$$

where $d^* = \text{ess sup}_P d \leq h \log \left( \frac{e(k + 1)N}{h} \right) + \log(\epsilon^{-1}).$
Choosing a fixed $\lambda$ and optimizing it at the end, we can also prove that

**Theorem.** For any $\zeta > 1$, for any $\epsilon \leq e^{-1}$, any integer $N \geq 4\zeta$, with $\mathbb{P}$ probability at least $1 - \epsilon$, for any $\theta \in \Theta$,

$$R(\theta) \leq r_1(\theta) + \frac{\zeta d}{N} + \sqrt{\frac{\zeta^2 d^2}{N^2} + \frac{2\zeta (1 + k^{-1}) r_1(\theta)}{N}},$$

where

$$d = \mathbb{P}\left\{ \log \left[ N(X_1^{(k+1)N}) \right] \mid Z_1^N \right\} + \log \left( e^{-1} \left( \frac{\log(2N)}{\log(k)} + 1 \right) \right) \geq 1$$
This is to be compared with Vapnik’s result

**Theorem (Vapnik).** With P probability at least $1 - \epsilon$,

$$R(\theta) \leq r_1(\theta) + \frac{2d'}{N} \left(1 + \sqrt{1 + \frac{N r_1(\theta)}{d'}}\right),$$

where $d' = \log \left\{P^\otimes 2^N \left[ N(X_1^{2N}) \right] \right\} + \log(4\epsilon^{-1})$. 

16
Instead of looking for an improved Vapnik’s bound, we can also optimize the right-hand side of the learning bound, leading to

**Theorem.** With $P$ probability at least $1 - \epsilon$,

$$
\hat{\rho}_{\lambda + \frac{\lambda^2}{2N}} [r_2(\theta)] \leq \left( \lambda - \frac{\lambda^2}{2N} \right)^{-1} \left\{ -\log \left[ \pi \left\{ \exp \left[ - \left( \lambda + \frac{\lambda^2}{2N} \right) r_1(\theta) \right] \right\} \right\} \\
+ \log(e^{-1}) \right\} \\
= \frac{1 + \frac{\lambda}{2N}}{1 - \frac{\lambda}{2N}} \left\{ \frac{1}{\lambda + \frac{\lambda^2}{2N}} \int_0^{\lambda + \frac{\lambda^2}{2N}} \hat{\rho}_\beta [r_1(\theta)] d\beta \right\} + \frac{\log(e^{-1})}{\lambda - \frac{\lambda^2}{2N}}
$$

where

$$
d\hat{\rho}_\beta(\theta) = \frac{\exp[-\beta r_1(\theta)]}{\pi \left\{ \exp[-\beta r_1(\theta)] \right\}} d\pi(\theta).
$$
Localization

We will restrict for simplicity to the case when $k = 1$ (i.e. the training set and test set have the same size). Let us put

$$
\eta(\theta) = \left( \frac{\lambda^2}{2N} + \beta \right) \left[ r_1(\theta) + r_2(\theta) \right]
+ \log \left\{ \pi \exp \left[ -\beta \left[ r_1(\theta) + r_2(\theta) \right] \right] \right\} + \log(\epsilon^{-1}),
$$

to get

18
Theorem. With $P$ probability at least $1 - \epsilon$, for any posterior probability measure $\rho \in \mathcal{M}_+^1$,

$$r_2(\theta) \leq \left\{ \left[ (1 - \xi)\lambda - (1 + \xi) \frac{\lambda^2}{2N} \right]^{-1} \left[ (1 - \xi)\lambda + (1 + \xi) \frac{\lambda^2}{2N} \right] \rho[r_1(\theta)] + \mathcal{K}(\rho, \hat{\rho}_{2\xi \lambda}) + (1 + \xi) \log(\frac{2}{\epsilon}) \right\}.$$
Corollary. With $P$ probability at least $1 - \epsilon$,

$$\hat{\rho}_{(1+\xi)\lambda(1+\frac{1}{2N})}[r_2(\theta)] \leq \left[ (1-\xi)\lambda -(1+\xi)\frac{\lambda^2}{2N} \right]^{-1} \left\{ \int_{2\xi\lambda}^{(1+\xi)\lambda(1+\frac{1}{2N})} \hat{\rho}_\beta [r_1(\theta)] d\beta + (1 + \xi) \log\left(\frac{2}{\epsilon}\right) \right\}$$

$$\leq \left[ (1-\xi)\lambda -(1+\xi)\frac{\lambda^2}{2N} \right]^{-1} \left\{ (1-\xi)\lambda + (1+\xi)\frac{\lambda^2}{2N} \right\} \hat{\rho}_{2\xi\lambda}[r_1(\theta)] + (1 + \xi) \log\left(\frac{2}{\epsilon}\right).$$

In the same way, with $P$ probability at least $1 - \epsilon$,

$$\hat{\rho}_\lambda[r_2(\theta)] \leq \left[ 1 + \frac{(1+\xi)\lambda}{4\xi(1-\xi)N} \right] \hat{\rho}_\lambda [r_1(\theta)] + \frac{2\xi(1+\xi)}{(1-\xi)\lambda} \log\left(\frac{2}{\epsilon}\right) \frac{1}{1 - \frac{(1+\xi)\lambda}{4\xi(1-\xi)N}}.$$
As a special case, choosing $\xi = 8^{-1/2}$ we get

$$
\hat{\rho}_\lambda[r_2(\theta)] \leq \frac{(1 + \frac{3\lambda}{2N}) \hat{\rho}_\lambda[r_1(\theta)] + \frac{3}{2N} \log\left(\frac{2}{\epsilon}\right)}{1 - \frac{3\lambda}{2N}}.
$$
Compression schemes

- Let us consider some estimator
\[ \hat{f} : \bigcup_{n=1}^{+\infty} (X \times Y)^n \times X \rightarrow Y; \]

- Let us put for any training set \( Z' = (x'_i, y'_i)_{i=1}^n \in (X \times Y) \)
\[ \hat{f}_{Z'}(x) = \hat{f}(Z', x) \quad x \in X. \]

- Let us assume that \( Z' \mapsto \hat{f}_{Z'} \) is an exchangeable function of \( Z' \).
For any given sample $Z = (X_i, Y_i)_{i=1}^{2N}$, let us consider the model

$$
\mathcal{R}_h = \{ \hat{f}(x'_i, y'_i)^h : \{x'_i : 1 \leq i \leq h\} \subset \{X_i : 1 \leq i \leq 2N\},
(\gamma'_i)^j \in \gamma^h \}.
$$

Let $\mathcal{R} = \bigcup_{h=1}^{N} \mathcal{R}_h$ be the disjoint union of these models.

Let $\pi \in \mathcal{M}_1^1(\mathcal{R})$ be a prior measure which is uniform on each $\mathcal{R}_h$ and such that for some given parameter $\alpha \in [0, 1[$

$$
\pi(\mathcal{R}_h) \geq (1 - \alpha)\alpha^h.
$$

It is easy to see that

$$
\log|\mathcal{R}_h| = \log\left(\binom{2N}{h} |\gamma|^h\right) \leq h \left[ \log\left(\frac{2N}{h}\right) + 1 + \log(|\gamma|) \right].
$$
Theorem. For any \( \alpha \in ]0,1[ \), any \( \zeta > 1 \), with \( \mathbb{P} \) probability at least \( 1 - \epsilon \), for any \( h = 1, \ldots , 2N \), any \( f \in \mathcal{R}_h \)
\[
r_2(f) \leq \inf_{\lambda \in [1,2N]} B(\lambda, h, f),
\]
where
\[
B(\lambda, h, f) = \left(1 - \frac{\zeta \lambda}{2N}\right)^{-1} \left\{ \left(1 + \frac{\zeta \lambda}{2N}\right) r_1(f) + \frac{1}{\lambda} \left[ -\log(1 - \alpha) + h \left[ \log\left(\frac{N}{\alpha}\right) + 1 + \log(|y|) - \log(\alpha) \right] + \log(\epsilon^{-1}) + \log \left( \frac{\log(2N)}{\log(\alpha)} + 1 \right) \right] \right\}.
\]
We can then build an adaptive estimator \( \hat{f}_a \) by minimizing \( B(\lambda, h, f) \).
Let \( \mathcal{R}_h \) be the observable part of \( \mathcal{R}_h \), more precisely, let us put
\[
\hat{\mathcal{R}}_h = \left\{ \hat{f}(x_i, y_i) : 1 \leq i \leq h \right\} \subset \left\{ X_i : 1 \leq i \leq N \right\}, (y_i^h)_{i=1}^h \in \mathcal{Y}_h \).
\]
Let us define
\[
\hat{h} = \arg \min_{h=1, \ldots, N} \inf \left\{ B(\lambda, h, f), \lambda \in [1, 2N], f \in \hat{\mathcal{R}}_h \right\}
\]
\[
\hat{f}_a = \arg \min_{f \in \hat{\mathcal{R}}_h} \inf_{\lambda \in [1, 2N]} B(\lambda, \hat{h}, f).
\]
**Proposition.** With these notations
\[
r_2(\hat{f}_a) \leq \inf \left\{ B(\lambda, h, f) : \lambda \in [1, 2N], h \in [1, N], f \in \hat{\mathcal{R}}_h \right\}.
\]
In the transductive case (i.e. when $X_{N+1}^N$ is observed), the exchangeable model $\mathcal{R}_h$ is observable, and therefore we can simulate the Gibbs posterior distribution (e.g. using some MCMC method) and compute localized learning bounds.

Natural applications of compression schemes are:
- bounding the generalization error of SVMs as a function of the number of support vectors;
- pruning decision trees, or even choosing the questions to ask at each node in some data driven way.
Margin bounds for SVMs

- Assume that \((X_i)_{i=1}^{2N}\) and \((Y_i)_{i=1}^{N}\) are observed;
- Let \(K\) be some symmetric positive kernel on \(X\);
- For any \(K\)-separable training set \(Z' = (X_i, y'_i)_{i=1}^{2N}\), where \((y'_i)_{i=1}^{2N} \in \{0,1\}^{2N}\), let us consider the SVM \(\hat{f}_{Z'}\) defined by \(K\) and \(Z'\).
Let \(\gamma(Z')\) be its margin.

Let \(R^2 = \max_{i=1,...,2N} K(x_i, x_i) \)
\[+ \frac{1}{4N^2} \sum_{j=1}^{2N} \sum_{k=1}^{2N} K(x_j, x_k) - \frac{1}{N} \sum_{j=1}^{2N} K(x_i, x_j).\]
For any integer $h = 1, \ldots, N$ let us define the margin values

$$\gamma_{2h} = \frac{R}{\sqrt{2h - 1}}$$

$$\gamma_{2h + 1} = \frac{R}{\sqrt{2h \left( 1 - \frac{1}{(2h + 1)^2} \right)}}$$

and the exchangeable model

$$\mathcal{R}_h = \{ f_{Z'} : Z' = (X_i, y_i')_{i=1}^{2N} \text{ is } K\text{-separable and } \gamma(Z') \geq \gamma_h \}.$$ 

The models $\mathcal{R}_h, h = 1, \ldots, N$ are nested, moreover

$$\log(|\mathcal{R}_h|) \leq h \left[ \log\left( \frac{2N}{K} \right) + 1 \right].$$
Proposition. For any $\alpha \in ]0, 1[$, any $\zeta > 1$, with $\mathbb{P}$ probability at least $1 - \epsilon$, for any $h = 1, \ldots, N$, any SVM $f \in \mathcal{R}_h$,

$$r_2(f) \leq \inf_{\lambda \in [1, 2N]} \left( 1 - \frac{\zeta \lambda}{2N} \right)^{-1} \left\{ \left( 1 + \frac{\zeta \lambda}{2N} \right) r_1(f) \right. \right.$$

$$\left. + \frac{1}{\lambda} \left[h \left[ \log\left( \frac{2N}{h} \right) + 1 - \log(\alpha) \right] - \log(1 - \alpha) \right. \right.$$

$$\left. - \log(\epsilon) + \log \left[ \log \frac{\log(2N)}{\log(\zeta)} + 1 \right] \right\}.$$
It is also possible to get bounds involving the margin on the training set (and not on the union of the training and test sets). This is based on a combinatorial lemma by Alon, Ben-David, Cesa-Bianchi and Haussler: Let $\mathcal{X} = \{1, \ldots, n\}$ and $\mathcal{Y} = \{1, \ldots, b\}$, where $b \geq 3$. Let $\mathcal{R} = \{f : \mathcal{X} \to \mathcal{Y}\}$ be some set of classification rules. A pair $(A, s)$ where $A \subset \mathcal{X}$ and $s : A \to \mathcal{Y}$ is said to be shattered by $\mathcal{R}$ if for any $(\sigma_x)_{x \in A} \in \{-1, +1\}^A$ there exists $f \in \mathcal{R}$ such that

$$\min_{x \in A} \sigma_x [f(x) - s(x)] \geq 1.$$
The fat shattering dimension of $\mathcal{R}$ is defined as the maximal size $|A|$ of pairs $(A, s)$ shattered by $\mathcal{R}$.

**Lemma.** As soon as this fat shattering dimension is not greater than $h$, there exists a 1-net $F$ for the norm $\mathcal{L}_\infty$ on $\mathcal{R}$ of size

$$\log(|F|) < \log((b-1)(b-2)n) \left\{ \frac{\log\left[ \sum_{i=1}^{h} \binom{n}{i}(b-2)^i \right]}{\log(2)} + 1 \right\} + \log(2)$$

$$\leq \log((b-1)(b-2)n) \left\{ \left[ \log\left( \frac{(b-2)n}{h} \right) + 1 \right] + \frac{h}{\log(2)} + 1 \right\} + \log(2).$$
Application to SVMs: it is enough to deal with the linear case.

- Let $X = \mathbb{R}^d$ et $\{ -1, +1 \}$
- Let $R \geq \max\{\|X_i\| : 1 \leq i \leq 2N\}$
- $\Theta = \{(w, b) \in \mathbb{R}^d \times \mathbb{R} : \|w\| = 1\}$
- $g_{w,b}(x) = \langle w, x \rangle - b$
- $G_{w,b}(x) = \text{sign} \left[ g_{w,b}(x) \right]$

**Theorem.** With $\mathbb{P}$ probability at least $1 - \epsilon$,

$$
\frac{1}{N} \sum_{i=N+1}^{2N} 1 \left[ G_{w,b}(X_i) \neq Y_i \right] \\
\leq \left( 1 - \frac{\lambda}{2N} \right)^{-1} \left\{ \left( 1 + \frac{\lambda}{2N} \right) \frac{1}{N} \sum_{i=1}^{N} 1 \left[ g_{w,b}(X_i) Y_i \leq 4\gamma_h \right] \\
+ \frac{1}{\lambda} \left\{ \log(40N) \left\{ \frac{h}{\log(2)} \log \left( \frac{8eN}{\lambda} \right) + 1 \right\} + \log(2^{\epsilon^{-1}}) \right\}\right\}
$$