

Learning, information theory and thermodynamics

Olivier Catoni

CNRS
Laboratoire de Probabilités et Modèles Aléatoires
Université Paris 6
catoni@ccr.jussieu.fr

Lundi 18 juin 2007 / Rennes

*This is a synthesis of contributions from the PhD works of two of my former students, **Jean-Yves Audibert** and **Pierre Alquier**, and of my own point of view on PAC-Bayesian learning.*

<http://cermics.enpc.fr/audibert>

<http://www.crest.fr/pageperso/alquier/alquier.htm>

<http://www.proba.jussieu.fr/users/catoni/homepage/homepage-en.html>

Empirical processes and risk minimization

A general framework.

Let $\theta \mapsto W_i(\theta)$, $\theta \in \Theta$, $i = 1, \dots, N$ be N **independent** real valued stochastic processes indexed by some measurable set Θ .

We will eventually (in the very end of this talk) focus on the **binary case** where $W_i(\theta) \in \{0, 1\}$, which corresponds to **classification**. Let us consider

$$\theta \mapsto r(\theta) = \frac{1}{N} \sum_{i=1}^N W_i(\theta), \quad \text{the empirical process,}$$

$$\theta \mapsto R(\theta) = \mathbf{E}[r(\theta)], \quad \text{its mean.}$$

We **do not** consider the normalized process

$\theta \mapsto \sqrt{N} [r(\theta) - R(\theta)]$, **because** we are interested in the following

Question

How can we estimate $\arg \min R$ from $\arg \min r$?

Empirical processes and risk minimization

A general framework.

Let $\theta \mapsto W_i(\theta)$, $\theta \in \Theta$, $i = 1, \dots, N$ be N **independent** real valued stochastic processes indexed by some measurable set Θ . We will eventually (in the very end of this talk) focus on the **binary case** where $W_i(\theta) \in \{0, 1\}$, which corresponds to **classification**. Let us consider

$$\theta \mapsto r(\theta) = \frac{1}{N} \sum_{i=1}^N W_i(\theta), \quad \text{the empirical process,}$$

$$\theta \mapsto R(\theta) = \mathbf{E}[r(\theta)], \quad \text{its mean.}$$

We **do not** consider the normalized process $\theta \mapsto \sqrt{N} [r(\theta) - R(\theta)]$, **because** we are interested in the following

Question

How can we estimate **$\arg \min R$** from **$\arg \min r$** ?

Supervised classification setting

We observe **independent labelled patterns**

$$(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y},$$

where \mathcal{X} is some big pattern space and \mathcal{Y} a finite set of labels.

We consider a family of classification rules

$$f_\theta : \mathcal{X} \rightarrow \mathcal{Y}, \theta \in \Theta,$$

and we define the **empirical error rate** by setting

$$W_i(\theta) = \mathbb{1}[f_\theta(X_i) \neq Y_i],$$

so that

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

We would like to minimize the **generalization error rate**

$$R(\theta) = \mathbf{E}[r(\theta)] = \frac{1}{N} \sum_{i=1}^N \mathbf{P}[f_\theta(X_i) \neq Y_i].$$

Confidence interval for an empirical mean

Here the parameter $\theta \in \Theta$ is fixed. First attempt : compute

$$\mathbf{E} \left\{ \exp[-\lambda r(\theta)] \right\} = \prod_{i=1}^N \mathbf{E} \left\{ \exp \left[-\frac{\lambda}{N} W_i(\theta) \right] \right\};$$

not bad, but yields a confidence interval for $r(\theta)$, not for $R(\theta)$.
Second attempt :

$$\begin{aligned} \mathbf{E} \left\{ \exp \left\{ \sum_{i=1}^N \log \left[1 - \frac{\lambda}{N} W_i(\theta) \right] \right\} \right\} &= \prod_{i=1}^N \left\{ 1 - \mathbf{E} \left[\frac{\lambda}{N} W_i(\theta) \right] \right\} \\ &\leq \exp \left\{ N \log \left[1 - \frac{\lambda}{N} R(\theta) \right] \right\}; \end{aligned}$$

much more useful !

Putting $\mathbb{P}_\theta = \frac{1}{N} \sum_{i=1}^N \delta_{W_i(\theta)}$, and $\Phi_a(w) = -a^{-1} \log(1 - aw)$, we get $\mathbf{E} \left\{ \exp \left[-\lambda \mathbb{P}_\theta \left(\Phi_{\frac{\lambda}{N}} \right) \right] \right\} \stackrel{\text{in the i.i.d. case}}{\leq} \exp \left\{ -\lambda \Phi_{\frac{\lambda}{N}} [R(\theta)] \right\}$, and thus

Theorem

For any $\lambda > 0$ such that $-\frac{N}{\lambda} \leq W_i(\theta) \leq \frac{N}{\lambda}$ a. s.,

$$\mathbf{P} \left[R(\theta) \leq \Phi_{\frac{\lambda}{N}}^{-1} \left(\mathbb{P}_\theta \left(\Phi_{\frac{\lambda}{N}} \right) - \frac{\log(\epsilon)}{\lambda} \right) \right] \geq 1 - \epsilon,$$

$$\mathbf{P} \left[R(\theta) \geq \Phi_{-\frac{\lambda}{N}}^{-1} \left(\mathbb{P}_\theta \left(\Phi_{-\frac{\lambda}{N}} \right) + \frac{\log(\epsilon)}{\lambda} \right) \right] \geq 1 - \epsilon.$$

Better than Hoeffding's, Bennett's or Bernstein's inequalities : $\mathbb{P}_\theta \left(\Phi_{\frac{\lambda}{N}} \right) = r(\theta) + \mathbb{P}_\theta \left(\Phi_{\frac{\lambda}{N}} - \text{Id} \right) \simeq r(\theta) + \frac{\lambda}{2N} \int_{\mathbb{R}} w^2 \mathbb{P}_\theta(dw)$, so that we use some kind of **empirical second moment estimate** straightaway.

... and what if $W_i(\theta)$ are not bounded ?

... you can **truncate** them, a little thinking shows that the order of magnitude of the truncation level $\frac{N}{\lambda}$ is not improvable,
... or you can **make another choice of Φ** , considering for some $\rho > 1$ and $b \in]0, 1[$

$$\Phi_a(x) = -a^{-1} \log \left[1 - aw + \left(\frac{\rho - 1}{1 - b} \right)^{\rho-1} \left(\frac{(aw)_+}{\rho} \right)^\rho \right] \\ \leq -a^{-1} \log(b), \quad x \in \mathbb{R}.$$

This gives for any $\lambda > 0$, with probability at least $1 - \epsilon$

$$R(\theta) \leq \frac{N}{\lambda} \left\{ 1 - \exp \left[-\frac{\lambda}{N} \left(\mathbb{P}_\theta(\Phi_{\frac{\lambda}{N}}) - \frac{\log(\epsilon)}{\lambda} \right) \right] \right\} \\ + \left[\frac{(\rho - 1)\lambda}{(1 - b)N} \right]^{\rho-1} \mathbf{E} \left\{ \mathbb{P}_\theta \left[\left(\frac{w_+}{\rho} \right)^\rho \right] \right\}.$$

... and of course also the reverse ...

With probability at least $1 - \epsilon$

$$R(\theta) \geq \frac{N}{\lambda} \left\{ \exp \left[\frac{\lambda}{N} \left(\mathbb{P}_\theta(\Phi_{-\frac{\lambda}{N}}) + \frac{\log(\epsilon)}{\lambda} \right) \right] - 1 \right\} \\ - \left[\frac{(\rho - 1)\lambda}{(1 - b)N} \right]^{\rho - 1} \mathbf{E} \left\{ \mathbb{P}_\theta \left[\left(\frac{w_-}{\rho} \right)^\rho \right] \right\}.$$

One model

Let $\hat{\theta} \in \arg \min_{\Theta} r$. What can we say about $R(\hat{\theta})$?

Statistical mechanics : at least for $\lambda > 0$ large enough, given some prior probability measure $\pi \in \mathcal{M}_+^1(\Theta)$,

$$\mathbf{E} \left[\exp(-\lambda r(\hat{\theta})) \right] \simeq \mathbf{E} \left[\pi \left(\exp(-\lambda r) \right) \right],$$

and introducing the **Gibbs posterior distribution**

$$\frac{d\pi_{\exp(-\lambda r)}}{d\pi}(\theta) = \frac{\exp(-\lambda r(\theta))}{\pi[\exp(-\lambda r)]},$$

$$r(\hat{\theta}) \simeq \pi_{\exp(-\lambda r)}(r).$$

This gives the idea of approximating $\hat{\theta}$ by a **randomized** $\tilde{\theta}$. In the following we will let **$\rho(d\theta | W)$ be the distribution of $\tilde{\theta}$ knowing the sample W** . (Here above we chose $\rho = \pi_{\exp(-\lambda r)} \cdot$)

Using the same trick again ...

$$\mathbf{E} \left\{ \pi \left[\exp \left\{ \lambda \left[\Phi_{\frac{\lambda}{N}}(R) - \mathbb{P}_{\theta}(\Phi_{\frac{\lambda}{N}}) \right] \right\} \right] \right\} \stackrel{\text{when } W \text{ i.i.d!}}{\leq} 1,$$

Fubini!

and the simple fact $\pi[\exp(h)] = \rho \left\{ \exp \left[h - \log \left(\frac{d\rho}{d\pi} \right) \right] \right\}$, we get

Theorem

For any posterior distribution $\rho : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$, with $\mathbf{P}\rho$ probability at least $1 - \epsilon$,

$$R(\tilde{\theta}) \leq \Phi_{\frac{\lambda}{N}}^{-1} \left\{ \mathbb{P}_{\tilde{\theta}}(\Phi_{\frac{\lambda}{N}}) + \frac{\log \left[\frac{d\rho}{d\pi}(\tilde{\theta}) \right] - \log(\epsilon)}{\lambda} \right\}.$$

Consequence : **you do not need to randomize much**, for instance if $\Theta = (0, 1)^d$, π is the Lebesgue measure and you draw $\tilde{\theta}$ at random in a box of edge N^{-1} containing $\hat{\theta}$, you get $\log \left[\frac{d\rho}{d\pi}(\tilde{\theta}) \right] = d \log(N)$.

Choice of the prior

Introducing the **Kullback divergence**

$$\mathcal{K}(\rho, \pi) = \begin{cases} \rho \left[\log \left(\frac{d\rho}{d\pi} \right) \right] & \text{if } \rho \ll \pi \\ +\infty & \text{otherwise} \end{cases},$$

we have on average

$$\mathbf{P}_\rho(R) \leq \Phi_{\frac{\lambda}{N}}^{-1} \left\{ \mathbf{P}_\rho[\mathbb{P}(\cdot | \Phi_{\frac{\lambda}{N}})] + \frac{\mathbf{P}[\mathcal{K}(\rho, \pi)] - \log(\epsilon)}{\lambda} \right\}.$$

↓
expected generalization
risk of $\tilde{\theta}$

↓
expected modified
empirical risk of $\hat{\theta}$.

As $\mathbf{P}[\mathcal{K}(\rho, \pi)] = \mathbf{P}\left\{ \mathcal{K}[\rho, \mathbf{P}(\rho)] \right\} + \mathcal{K}[\mathbf{P}(\rho), \pi]$, **the most desirable** choice of π would be $\pi = \mathbf{P}(\rho)$. Moreover $\mathbf{P}\left\{ \mathcal{K}[\rho, \mathbf{P}(\rho)] \right\}$ is the **mutual information between $\tilde{\theta}$ and W** .

Localization

But for a fixed π , the previous bound is optimized by

$$\rho = \pi_{\exp[-\lambda \mathbb{P}(\cdot | \Phi_{\frac{\lambda}{N}})]} \simeq \pi_{\exp(-\lambda r)},$$

leading to the idea of **localization** : work with ρ close to $\pi_{\exp(-\beta r)}$ and choose as a prior $\pi_{\exp(-\lambda R)}$.

Requirement :

to find an **empirical bound for $\mathcal{K}[\rho, \pi_{\exp(-\beta R)}]$** .

Solution :

$$\begin{aligned} \text{Write } \mathcal{K}[\rho, \pi_{\exp(-\beta R)}] &= \beta [\rho(R) - \pi_{\exp(-\beta R)}(R)] \\ &\quad + \mathcal{K}(\rho, \pi) - \mathcal{K}[\pi_{\exp(-\beta R)}, \pi], \end{aligned}$$

bound $\rho(R) - \pi_{\exp(-\beta R)}(R)$ by $\rho(r) - \pi_{\exp(-\beta R)}(r)$ and replace $\pi_{\exp(-\beta R)}$ by a supremum over all possible posterior distributions.

Theorem

With probability at least $1 - \epsilon$, for any posterior distribution

$$\rho : \Omega \rightarrow \mathcal{M}_+^1(\Theta),$$

$$\begin{aligned} \mathcal{K}(\rho, \pi_{\exp(-\beta R)}) &\leq \left(1 - \frac{\beta}{\lambda}\right)^{-1} \left\{ \mathcal{K}[\rho, \pi_{\exp(-\beta r)}] \right. \\ &+ \int \rho(d\theta_1) \log \left\{ \int \pi_{\exp(-\beta r)}(d\theta_2) \right. \\ &\quad \times \exp \left[\beta \int \mathbb{P}_{\theta_1, \theta_2}(dw_1, dw_2) (\Phi_{\frac{\lambda}{N}} - \text{Id})(w_1 - w_2) \right] \left. \right\} \\ &\quad \left. - \frac{\beta}{\lambda} \log(\epsilon) \right\}, \end{aligned}$$

$$\text{where } \mathbb{P}_{\theta_1, \theta_2} = \frac{1}{N} \sum_{i=1}^N \delta_{[w_i(\theta_1), w_i(\theta_2)]}.$$

Disintegrated variant

Theorem

For any posterior distribution $\rho : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$, with \mathbf{P}_ρ probability at least $1 - \epsilon$,

$$\begin{aligned} \log\left(\frac{d\rho}{d\pi_{\exp(-\beta R)}}(\tilde{\theta})\right) &\leq \left(1 - \frac{\beta}{\lambda}\right)^{-1} \left\{ \log\left(\frac{d\rho}{d\pi_{\exp(-\beta r)}}(\tilde{\theta})\right) \right. \\ &+ \log\left\{ \int \pi_{\exp(-\beta r)}(d\theta) \exp\left[\frac{\beta}{N} \sum_{i=1}^N (\Phi_{\frac{\lambda}{N}} - \text{Id}) [W_i(\tilde{\theta}) - W_i(\theta)]\right] \right\} \\ &\quad \left. - \frac{\beta}{\lambda} \log(\epsilon) \right\}. \end{aligned}$$

Theorem

For any posterior $\rho : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$, with \mathbf{P}_ρ probability at least $1 - \epsilon$,

$$\begin{aligned} \Phi_{\frac{\alpha}{N}}[R(\tilde{\theta})] &\leq \mathbb{P}_{\tilde{\theta}}[\Phi_{\frac{\alpha}{N}}] \\ &+ \frac{\lambda}{\alpha(\lambda - \beta)} \left\{ \log \left[\int \pi_{\exp(-\beta r)}(d\theta) \right. \right. \\ &\quad \times \exp \left\{ \beta \int \mathbb{P}_{\tilde{\theta}, \theta}(dw_1, dw_2) (\Phi_{\frac{\lambda}{N}} - \text{Id})(w_1 - w_2) \right\} \left. \right] \\ &\quad \left. + \log \left[\frac{d\rho}{d\pi_{\exp(-\beta r)}}(\tilde{\theta}) \right] - \log \left(\frac{\epsilon}{2} \right) \right\}. \end{aligned}$$

Recall that $(\Phi_{\frac{\lambda}{N}} - \text{Id})(w_1 - w_2) \simeq \frac{\lambda}{2N}(w_1 - w_2)^2$ and think about taking $\lambda = 2\beta$.

Model selection

(and parameter estimation)

Let us assume that the parameter set is decomposed into a **family of models** :

$$\Theta = \bigcup_{m \in M} \Theta_m.$$

Consider in each model some **empirical risk minimizer**

$$\hat{\theta}_m \in \arg \min_{\Theta_m} r,$$

and ask the **question**: Can we pick up some $\hat{\theta}_m$ with a **close to minimum generalization risk** $R(\hat{\theta}_m)$?

Model selection

(and parameter estimation)

Let us assume that the parameter set is decomposed into a **family of models** :

$$\Theta = \bigcup_{m \in M} \Theta_m.$$

Consider in each model some **empirical risk minimizer**

$$\hat{\theta}_m \in \arg \min_{\Theta_m} r,$$

and ask the **question**: Can we pick up some $\hat{\theta}_m$ with a **close to minimum generalization risk** $R(\hat{\theta}_m)$?

More generally, we can
replace Θ with $\Theta \times (M \sqcup \{o\})$, considering
 $[W_i(\theta, m)]_{(\theta, m) \in \Theta \times (M \sqcup \{o\})}$,

define $\hat{\theta}_m \in \arg \min_{\theta \in \Theta} r(\theta, m)$, for any $m \in M$,
and look for an estimate of $\arg \min_{m \in M} R(\hat{\theta}_m, o)$.

This covers the choice of a penalty, when
 $W_i(\theta, m) = W_i(\theta) + \gamma(m, \theta)$.

Let $\tilde{\theta}_m$ be a randomized approximation of $\hat{\theta}_m$, and let
 $\rho_m : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$ be the posterior distribution of $\tilde{\theta}_m$ knowing the
sample W .

The best control of $\arg \min_{m \in M} R(\tilde{\theta}_m, o)$ is provided by relative
bounds, that is confidence intervals for the differences
 $R(\tilde{\theta}_m, o) - R(\tilde{\theta}_{m'}, o)$.

More generally, we can

replace Θ with $\Theta \times (M \sqcup \{o\})$, considering

$$[W_i(\theta, m)]_{(\theta, m) \in \Theta \times (M \sqcup \{o\})},$$

define $\hat{\theta}_m \in \arg \min_{\theta \in \Theta} r(\theta, m)$, for any $m \in M$,

and look for an estimate of $\arg \min_{m \in M} R(\hat{\theta}_m, o)$.

This covers the choice of a penalty, when

$$W_i(\theta, m) = W_i(\theta) + \gamma(m, \theta).$$

Let $\tilde{\theta}_m$ be a randomized approximation of $\hat{\theta}_m$, and let

$\rho_m : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$ be the posterior distribution of $\tilde{\theta}_m$ knowing the sample W .

The best control of $\arg \min_{m \in M} R(\tilde{\theta}_m, o)$ is provided by relative bounds, that is confidence intervals for the differences

$$R(\tilde{\theta}_m, o) - R(\tilde{\theta}_{m'}, o).$$

More generally, we can replace Θ with $\Theta \times (M \sqcup \{o\})$, considering $[W_i(\theta, m)]_{(\theta, m) \in \Theta \times (M \sqcup \{o\})}$, define $\hat{\theta}_m \in \arg \min_{\theta \in \Theta} r(\theta, m)$, for any $m \in M$, and look for an estimate of $\arg \min_{m \in M} R(\hat{\theta}_m, o)$.

This covers the choice of a penalty, when $W_i(\theta, m) = W_i(\theta) + \gamma(m, \theta)$.

Let $\tilde{\theta}_m$ be a randomized approximation of $\hat{\theta}_m$, and let $\rho_m : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$ be the posterior distribution of $\tilde{\theta}_m$ knowing the sample W .

The best control of $\arg \min_{m \in M} R(\tilde{\theta}_m, o)$ is provided by relative bounds, that is confidence intervals for the differences $R(\tilde{\theta}_m, o) - R(\tilde{\theta}_{m'}, o)$.

More generally, we can replace Θ with $\Theta \times (M \sqcup \{o\})$, considering $[W_i(\theta, m)]_{(\theta, m) \in \Theta \times (M \sqcup \{o\})}$, define $\hat{\theta}_m \in \arg \min_{\theta \in \Theta} r(\theta, m)$, for any $m \in M$, and look for an estimate of $\arg \min_{m \in M} R(\hat{\theta}_m, o)$. This covers the **choice of a penalty**, when $W_i(\theta, m) = W_i(\theta) + \gamma(m, \theta)$.

Let $\tilde{\theta}_m$ be a **randomized approximation** of $\hat{\theta}_m$, and let $\rho_m : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$ be the posterior distribution of $\tilde{\theta}_m$ knowing the sample W .

The best control of $\arg \min_{m \in M} R(\tilde{\theta}_m, o)$ is provided by **relative bounds**, that is confidence intervals for the differences $R(\tilde{\theta}_m, o) - R(\tilde{\theta}_{m'}, o)$.

More generally, we can replace Θ with $\Theta \times (M \sqcup \{o\})$, considering $[W_i(\theta, m)]_{(\theta, m) \in \Theta \times (M \sqcup \{o\})}$, define $\hat{\theta}_m \in \arg \min_{\theta \in \Theta} r(\theta, m)$, for any $m \in M$, and look for an estimate of $\arg \min_{m \in M} R(\hat{\theta}_m, o)$. This covers the choice of a penalty, when $W_i(\theta, m) = W_i(\theta) + \gamma(m, \theta)$.

Let $\tilde{\theta}_m$ be a randomized approximation of $\hat{\theta}_m$, and let $\rho_m : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$ be the posterior distribution of $\tilde{\theta}_m$ knowing the sample W .

The best control of $\arg \min_{m \in M} R(\tilde{\theta}_m, o)$ is provided by relative bounds, that is confidence intervals for the differences $R(\tilde{\theta}_m, o) - R(\tilde{\theta}_{m'}, o)$.

More generally, we can replace Θ with $\Theta \times (M \sqcup \{o\})$, considering $[W_i(\theta, m)]_{(\theta, m) \in \Theta \times (M \sqcup \{o\})}$, define $\hat{\theta}_m \in \arg \min_{\theta \in \Theta} r(\theta, m)$, for any $m \in M$, and look for an estimate of $\arg \min_{m \in M} R(\hat{\theta}_m, o)$. This covers the choice of a penalty, when $W_i(\theta, m) = W_i(\theta) + \gamma(m, \theta)$.

Let $\tilde{\theta}_m$ be a randomized approximation of $\hat{\theta}_m$, and let $\rho_m : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$ be the posterior distribution of $\tilde{\theta}_m$ knowing the sample W .

The best control of $\arg \min_{m \in M} R(\tilde{\theta}_m, o)$ is provided by relative bounds, that is confidence intervals for the differences $R(\tilde{\theta}_m, o) - R(\tilde{\theta}_{m'}, o)$.

Consider $\pi \in \mathcal{M}_+^1(\Theta \times M)$, some atomic $\nu \in \mathcal{M}_+^1(\mathbb{R}_+)$, and a collection $\rho_{m,\beta} : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$, $m \in M$, $\beta \in \text{supp } \nu$ of posterior distributions.

Given the sample $\{W_i(\theta, m); \theta \in \Theta, m \in M \sqcup \{o\}, i = 1, \dots, N\}$, let us draw independently $\tilde{\theta}_{m,\beta}$, $m \in M$, $\beta \in \text{supp}(\nu)$, according to $\mathbf{P} \prod_{m,\beta} \rho_{m,\beta}$. Let

$$\begin{aligned} C(m, \beta) = & \inf_{\lambda \in \mathbb{R}_+} \frac{\lambda}{(\lambda - \beta)} \left\{ \log \left[\frac{d\rho_{m,\beta}}{d\pi_{\exp[-\beta r(\cdot, m)]}(\cdot | m)}(\tilde{\theta}_m) \right] \right. \\ & + \log \left[\int \pi_{\exp[-\beta r(\cdot, m)]}(d\theta | m) \right. \\ & \left. \left. \exp \left\{ \frac{\beta}{N} \sum_i (\Phi_{\frac{\lambda}{N}} - \text{Id}) [W_i(\tilde{\theta}_{m,\beta}, m) - W_i(\theta, m)] \right\} \right] \right. \\ & \left. \left. - \frac{\beta}{\lambda} \log \left(\frac{\epsilon \pi(m) \nu(\lambda) \nu(\beta)}{3} \right) \right\}. \end{aligned}$$

Lemma With probability at least $1 - \frac{\epsilon}{3}$, for any $m \in M$, $\beta \in \text{supp } \nu$,

$$\log \left[\frac{d\rho_{m,\beta}}{d\pi_{\exp[-\beta R(\cdot, m)]}(\cdot | m)}(\tilde{\theta}_{m,\beta}, m) \right] \leq C(m, \beta)$$

Theorem With $\mathbf{P} \prod_{m,\beta} \rho_{m,\beta}$ probability at least $1 - \epsilon$, for any $m, m' \in M$, $\alpha, \beta, \beta' \in \mathbb{R}_+$,

$$\begin{aligned} & \Phi_{\frac{\alpha}{N}} [R(\tilde{\theta}_{m',\beta'}, \mathbf{o}) - R(\tilde{\theta}_{m,\beta}, \mathbf{o})] \\ & \leq \frac{1}{N} \sum_{i=1}^N \Phi_{\frac{\alpha}{N}} [W_i(\tilde{\theta}_{m',\beta'}, \mathbf{o}) - W_i(\tilde{\theta}_{m,\beta}, \mathbf{o})] \\ & + \frac{1}{\alpha} \left\{ C(m, \beta) + C(m', \beta') - \log \left[\frac{\epsilon \pi(m) \pi(m') \nu(\alpha)}{3} \right] \right\}. \end{aligned}$$

Empirical optimization

Question : Assuming as previously proved that *with probability at least $1 - \epsilon$, for any $t, t' \in T$,*

$$R_{t'} - R_t \leq B(t, t'),$$

(where we put $t = (m, \beta)$,) how can we build an approximate minimizer of R_t ?

Remark : These bounds provide a **confidence interval of length $L(t, t') = B(t, t') + B(t', t)$** for $R_{t'} - R_t$, since by symmetry

$$-B(t', t) \leq R_{t'} - R_t \leq B(t, t').$$

Therefore it should be kept in mind that

$L(t, t') = B(t, t') + B(t', t)$ should be small (at least for N large enough).

Empirical optimization

Question : Assuming as previously proved that *with probability at least $1 - \epsilon$, for any $t, t' \in T$,*

$$R_{t'} - R_t \leq B(t, t'),$$

(where we put $t = (m, \beta)$,) how can we build an approximate minimizer of R_t ?

Remark : These bounds provide a **confidence interval of length $L(t, t') = B(t, t') + B(t', t)$** for $R_{t'} - R_t$, since by symmetry

$$-B(t', t) \leq R_{t'} - R_t \leq B(t, t').$$

Therefore it should be kept in mind that

$L(t, t') = B(t, t') + B(t', t)$ should be small (at least for N large enough).

Solution : consider w.l.o.g. that $t \in \mathcal{T} = \{1, \dots, T\}$ takes a finite number of values. Consider some arbitrary complexity function $(C_t)_{t \in \mathcal{T}}$ (we will apply the method with $C_t = C(m, \beta)$).

- ▶ Index R_t by **increasing complexity** $C_{t+1} \geq C_t$.
- ▶ **Chain the bound**, to make it subadditive, defining

$$\bar{B}(t, t') = \min \left\{ \sum_{i=1}^n B(s_{n-1}, s_n); \right. \\ \left. n \in \mathbb{N}, (s_i)_{i=0}^n \in \mathcal{T}^{n+1}, s_0 = t, s_n = t' \right\}$$

- ▶ Build an **empirical scale of performance** considering $p(t) = \min\{s : \bar{B}(s, t) > 0\}$ (the first index t is not proved to beat).
- ▶ Choose $\hat{t} = \min(\arg \max p)$.

Solution : consider w.l.o.g. that $t \in \mathcal{T} = \{1, \dots, T\}$ takes a finite number of values. Consider some arbitrary complexity function $(C_t)_{t \in \mathcal{T}}$ (we will apply the method with $C_t = C(m, \beta)$).

- ▶ Index R_t by **increasing complexity** $C_{t+1} \geq C_t$.
- ▶ **Chain the bound**, to make it subadditive, defining

$$\bar{B}(t, t') = \min \left\{ \sum_{i=1}^n B(s_{n-1}, s_n); \right. \\ \left. n \in \mathbb{N}, (s_i)_{i=0}^n \in \mathcal{T}^{n+1}, s_0 = t, s_n = t' \right\}$$

- ▶ Build an **empirical scale of performance** considering $p(t) = \min\{s : \bar{B}(s, t) > 0\}$ (the first index t is not proved to beat).
- ▶ Choose $\hat{t} = \min(\arg \max p)$.

Solution : consider w.l.o.g. that $t \in \mathcal{T} = \{1, \dots, T\}$ takes a finite number of values. Consider some arbitrary complexity function $(C_t)_{t \in \mathcal{T}}$ (we will apply the method with $C_t = C(m, \beta)$).

- ▶ Index R_t by **increasing complexity** $C_{t+1} \geq C_t$.
- ▶ **Chain the bound**, to make it subadditive, defining

$$\bar{B}(t, t') = \min \left\{ \sum_{i=1}^n B(s_{n-1}, s_n); \right. \\ \left. n \in \mathbb{N}, (s_i)_{i=0}^n \in \mathcal{T}^{n+1}, s_0 = t, s_n = t' \right\}$$

- ▶ Build an **empirical scale of performance** considering $p(t) = \min\{s : \bar{B}(s, t) > 0\}$ (the first index t is not proved to beat).
- ▶ Choose $\hat{t} = \min(\arg \max p)$.

Solution : consider w.l.o.g. that $t \in \mathcal{T} = \{1, \dots, T\}$ takes a finite number of values. Consider some arbitrary complexity function $(C_t)_{t \in \mathcal{T}}$ (we will apply the method with $C_t = C(m, \beta)$).

- ▶ Index R_t by **increasing complexity** $C_{t+1} \geq C_t$.
- ▶ **Chain the bound**, to make it subadditive, defining

$$\bar{B}(t, t') = \min \left\{ \sum_{i=1}^n B(s_{n-1}, s_n); \right. \\ \left. n \in \mathbb{N}, (s_i)_{i=0}^n \in \mathcal{T}^{n+1}, s_0 = t, s_n = t' \right\}$$

- ▶ Build an **empirical scale of performance** considering $p(t) = \min\{s : \bar{B}(s, t) > 0\}$ (the first index t is not proved to beat).
- ▶ Choose $\hat{t} = \min(\arg \max p)$.

Solution : consider w.l.o.g. that $t \in \mathcal{T} = \{1, \dots, T\}$ takes a finite number of values. Consider some arbitrary complexity function $(C_t)_{t \in \mathcal{T}}$ (we will apply the method with $C_t = C(m, \beta)$).

- ▶ Index R_t by **increasing complexity** $C_{t+1} \geq C_t$.
- ▶ **Chain the bound**, to make it subadditive, defining

$$\bar{B}(t, t') = \min \left\{ \sum_{i=1}^n B(s_{n-1}, s_n); \right. \\ \left. n \in \mathbb{N}, (s_i)_{i=0}^n \in \mathcal{T}^{n+1}, s_0 = t, s_n = t' \right\}$$

- ▶ Build an **empirical scale of performance** considering $p(t) = \min\{s : \bar{B}(s, t) > 0\}$ (the first index t is not proved to beat).
- ▶ Choose $\hat{t} = \min(\arg \max p)$.

Theorem : Let $\hat{p} = p(\hat{t})$. With probability at least $1 - \epsilon$,

$$\bar{B}(s, \hat{t}) \leq \begin{cases} 0, & 1 \leq s < \hat{p}, \\ \min_{s' < \hat{p}} B(s, s'), & \hat{p} \leq s < \hat{t}, \\ B(s, \hat{p}) + B(\hat{p}, \hat{t}), & s \in (\arg \max p), \\ B(s, \hat{t}), & s > \hat{p}, s \notin (\arg \max p). \end{cases}$$

Moreover

- ▶ when $\hat{p} \leq s < \hat{t}$, $s \notin (\arg \max p)$: there is $s' < \hat{p}$ such that $B(s', s) \geq \bar{B}(s', s) > 0$, thus $B(s, s') < L(s, s')$ and therefore $C_{s'} \leq C_s$ and $R_{\hat{t}} \leq R_{s'} \leq R_s + L(s, s')$;
- ▶ when $s \in (\arg \max p)$, $B(\hat{p}, s) > 0$ (by definition) and $B(\hat{t}, \hat{p}) > 0$ (requires chaining!). Thus $R_{\hat{t}} \leq R_s + L(s, \hat{p}) + L(\hat{p}, \hat{t})$. Meanwhile, $C_{\hat{p}} \leq C_s$ and $C_{\hat{t}} \leq C_s$, and $R_{\hat{p}} \leq R_s + L(s, \hat{p})$;
- ▶ when $s > \hat{t}$, $s \notin (\arg \max p)$, $C_{\hat{t}} \leq C_s$, using chaining: $B(\hat{t}, s) \geq \bar{B}(\hat{t}, s) \geq \bar{B}(s', s) - \bar{B}(s', \hat{t}) > 0$, for some $s' < \hat{p}$, thus $R_{\hat{t}} \leq R_s + L(s, \hat{t})$.

Theorem : Let $\hat{p} = p(\hat{t})$. With probability at least $1 - \epsilon$,

$$\bar{B}(s, \hat{t}) \leq \begin{cases} 0, & 1 \leq s < \hat{p}, \\ \min_{s' < \hat{p}} B(s, s'), & \hat{p} \leq s < \hat{t}, \\ B(s, \hat{p}) + B(\hat{p}, \hat{t}), & s \in (\arg \max p), \\ B(s, \hat{t}), & s > \hat{p}, s \notin (\arg \max p). \end{cases}$$

Moreover

- ▶ when $\hat{p} \leq s < \hat{t}$, $s \notin (\arg \max p)$: there is $s' < \hat{p}$ such that $B(s', s) \geq \bar{B}(s', s) > 0$, thus $B(s, s') < L(s, s')$ and therefore $C_{s'} \leq C_s$ and $R_{\hat{t}} \leq R_{s'} \leq R_s + L(s, s')$;
- ▶ when $s \in (\arg \max p)$, $B(\hat{p}, s) > 0$ (by definition) and $B(\hat{t}, \hat{p}) > 0$ (requires chaining!). Thus $R_{\hat{t}} \leq R_s + L(s, \hat{p}) + L(\hat{p}, \hat{t})$. Meanwhile, $C_{\hat{p}} \leq C_s$ and $C_{\hat{t}} \leq C_s$, and $R_{\hat{p}} \leq R_s + L(s, \hat{p})$;
- ▶ when $s > \hat{t}$, $s \notin (\arg \max p)$, $C_{\hat{t}} \leq C_s$, using chaining: $B(\hat{t}, s) \geq \bar{B}(\hat{t}, s) \geq \bar{B}(s', s) - \bar{B}(s', \hat{t}) > 0$, for some $s' < \hat{p}$, thus $R_{\hat{t}} \leq R_s + L(s, \hat{t})$.

Theorem : Let $\hat{p} = p(\hat{t})$. With probability at least $1 - \epsilon$,

$$\bar{B}(s, \hat{t}) \leq \begin{cases} 0, & 1 \leq s < \hat{p}, \\ \min_{s' < \hat{p}} B(s, s'), & \hat{p} \leq s < \hat{t}, \\ B(s, \hat{p}) + B(\hat{p}, \hat{t}), & s \in (\arg \max p), \\ B(s, \hat{t}), & s > \hat{p}, s \notin (\arg \max p). \end{cases}$$

Moreover

- ▶ when $\hat{p} \leq s < \hat{t}$, $s \notin (\arg \max p)$: there is $s' < \hat{p}$ such that $B(s', s) \geq \bar{B}(s', s) > 0$, thus $B(s, s') < L(s, s')$ and therefore $C_{s'} \leq C_s$ and $R_{\hat{t}} \leq R_{s'} \leq R_s + L(s, s')$;
- ▶ when $s \in (\arg \max p)$, $B(\hat{p}, s) > 0$ (by definition) and $B(\hat{t}, \hat{p}) > 0$ (requires chaining!). Thus $R_{\hat{t}} \leq R_s + L(s, \hat{p}) + L(\hat{p}, \hat{t})$. Meanwhile, $C_{\hat{p}} \leq C_s$ and $C_{\hat{t}} \leq C_s$, and $R_{\hat{p}} \leq R_s + L(s, \hat{p})$;
- ▶ when $s > \hat{t}$, $s \notin (\arg \max p)$, $C_{\hat{t}} \leq C_s$, using chaining: $B(\hat{t}, s) \geq \bar{B}(\hat{t}, s) \geq \bar{B}(s', s) - \bar{B}(s', \hat{t}) > 0$, for some $s' < \hat{p}$, thus $R_{\hat{t}} \leq R_s + L(s, \hat{t})$.

Theorem : Let $\hat{p} = p(\hat{t})$. With probability at least $1 - \epsilon$,

$$\bar{B}(s, \hat{t}) \leq \begin{cases} 0, & 1 \leq s < \hat{p}, \\ \min_{s' < \hat{p}} B(s, s'), & \hat{p} \leq s < \hat{t}, \\ B(s, \hat{p}) + B(\hat{p}, \hat{t}), & s \in (\arg \max p), \\ B(s, \hat{t}), & s > \hat{p}, s \notin (\arg \max p). \end{cases}$$

Moreover

- ▶ when $\hat{p} \leq s < \hat{t}$, $s \notin (\arg \max p)$: there is $s' < \hat{p}$ such that $B(s', s) \geq \bar{B}(s', s) > 0$, thus $B(s, s') < L(s, s')$ and therefore $C_{s'} \leq C_s$ and $R_{\hat{t}} \leq R_{s'} \leq R_s + L(s, s')$;
- ▶ when $s \in (\arg \max p)$, $B(\hat{p}, s) > 0$ (by definition) and $B(\hat{t}, \hat{p}) > 0$ (requires chaining!). Thus $R_{\hat{t}} \leq R_s + L(s, \hat{p}) + L(\hat{p}, \hat{t})$. Meanwhile, $C_{\hat{p}} \leq C_s$ and $C_{\hat{t}} \leq C_s$, and $R_{\hat{p}} \leq R_s + L(s, \hat{p})$;
- ▶ when $s > \hat{t}$, $s \notin (\arg \max p)$, $C_{\hat{t}} \leq C_s$, using chaining: $B(\hat{t}, s) \geq \bar{B}(\hat{t}, s) \geq \bar{B}(s', s) - \bar{B}(s', \hat{t}) > 0$, for some $s' < \hat{p}$, thus $R_{\hat{t}} \leq R_s + L(s, \hat{t})$.

Adaptive classification

$$\mathbb{1}[f_\theta(X_i) \neq Y_i] = W_i(\theta, o) \in \{0, 1\},$$

$$W_i(\theta, m) = \begin{cases} W_i(\theta, o), & \theta \in \Theta_m, \\ +\infty, & \text{otherwise.} \end{cases}$$

Let $R(\bar{\theta}, o) = \inf_{\theta \in \Theta} R(\theta, o)$, and

$$V(\theta, \bar{\theta}) = \mathbf{E} \left(\frac{1}{N} \sum_{i=1}^N |W_i(\theta, o) - W_i(\bar{\theta}, o)| \right) \\ \leq \mathbf{E} \left(\frac{1}{N} \sum_{i=1}^N \mathbb{1}[f_\theta(X_i) \neq f_{\bar{\theta}}(X_i)] \right).$$

Remark : Whenever $x \in \{-1, 0, 1\}$,

$$\Phi_a(x) - x = \frac{1}{2}(|x| + x)[\Phi_a(1) - 1] + \frac{1}{2}(|x| - 1)[\Phi(-1) + 1].$$

Let us make the **margin assumption**

$R(\theta, o) - R(\bar{\theta}, o) \geq c[V(\theta, \bar{\theta})]^\kappa$, $\theta \in \Theta$, and the **parametric complexity assumption**

$$\sup_{\beta \in \mathbb{R}_+} \beta \left[\int \pi_{\exp[-\beta R(\cdot, m)]}(d\theta, m) R(\theta, m) - \inf_{\theta \in \Theta} R(\theta, m) \right] \leq d_m.$$

Adaptive classification

$$\mathbb{1}[f_\theta(X_i) \neq Y_i] = W_i(\theta, o) \in \{0, 1\},$$

$$W_i(\theta, m) = \begin{cases} W_i(\theta, o), & \theta \in \Theta_m, \\ +\infty, & \text{otherwise.} \end{cases}$$

Let $R(\bar{\theta}, o) = \inf_{\theta \in \Theta} R(\theta, o)$, and

$$V(\theta, \bar{\theta}) = \mathbf{E} \left(\frac{1}{N} \sum_{i=1}^N |W_i(\theta, o) - W_i(\bar{\theta}, o)| \right) \\ \leq \mathbf{E} \left(\frac{1}{N} \sum_{i=1}^N \mathbb{1}[f_\theta(X_i) \neq f_{\bar{\theta}}(X_i)] \right).$$

Remark : Whenever $x \in \{-1, 0, 1\}$,

$$\Phi_a(x) - x = \frac{1}{2}(|x| + x)[\Phi_a(1) - 1] + \frac{1}{2}(|x| - 1)[\Phi(-1) + 1].$$

Let us make the **margin assumption**

$R(\theta, o) - R(\bar{\theta}, o) \geq c[V(\theta, \bar{\theta})]^\kappa$, $\theta \in \Theta$, and the **parametric complexity assumption**

$$\sup_{\beta \in \mathbb{R}_+} \beta \left[\int \pi_{\exp[-\beta R(\cdot, m)]}(d\theta, m) R(\theta, m) - \inf_{\theta \in \Theta} R(\theta, m) \right] \leq d_m.$$

Adaptive classification

$$\mathbb{1}[f_\theta(X_i) \neq Y_i] = W_i(\theta, o) \in \{0, 1\},$$

$$W_i(\theta, m) = \begin{cases} W_i(\theta, o), & \theta \in \Theta_m, \\ +\infty, & \text{otherwise.} \end{cases}$$

Let $R(\bar{\theta}, o) = \inf_{\theta \in \Theta} R(\theta, o)$, and

$$V(\theta, \bar{\theta}) = \mathbf{E} \left(\frac{1}{N} \sum_{i=1}^N |W_i(\theta, o) - W_i(\bar{\theta}, o)| \right) \\ \leq \mathbf{E} \left(\frac{1}{N} \sum_{i=1}^N \mathbb{1}[f_\theta(X_i) \neq f_{\bar{\theta}}(X_i)] \right).$$

Remark : Whenever $x \in \{-1, 0, 1\}$,

$$\Phi_a(x) - x = \frac{1}{2}(|x| + x)[\Phi_a(1) - 1] + \frac{1}{2}(|x| - 1)[\Phi(-1) + 1].$$

Let us make the **margin assumption**

$R(\theta, o) - R(\bar{\theta}, o) \geq c[V(\theta, \bar{\theta})]^k$, $\theta \in \Theta$, and the **parametric complexity assumption**

$$\sup_{\beta \in \mathbb{R}_+} \beta \left[\int \pi_{\exp[-\beta R(\cdot, m)]}(d\theta, m) R(\theta, m) - \inf_{\theta \in \Theta} R(\theta, m) \right] \leq d_m.$$

Theorem With probability at least $1 - \epsilon$,

$$\int \rho_{\widehat{m}, \widehat{\beta}}(d\theta) R(\theta, m) \leq \inf_{\theta \in \Theta} R(\theta, m) + \max \left\{ 847 C^{\frac{3}{2}} b^{\frac{\kappa-1}{2\kappa}} \left[\inf_{\theta \in \Theta} R(\theta, m) - \inf_{\theta \in \Theta} R(\theta, o) \right]^{\frac{1}{2\kappa}} \times \sqrt{\frac{d_m + \log\left(\frac{1 + \log_2(N)}{\epsilon \pi(m)}\right) + 5}{N}}, 2C [1082b]^{\frac{\kappa-1}{2\kappa-1}} 4^{\frac{1}{2\kappa-1}} \left\{ \frac{166C \left[d_m + \log\left(\frac{1 + \log_2(N)}{\epsilon \pi(m)}\right) + 5 \right]}{N} \right\}^{\frac{\kappa}{2\kappa-1}} \right\},$$

where $1 \simeq C \leq 3.2$, $b = (1 - \frac{1}{\kappa})(\kappa C)^{-\frac{1}{\kappa-1}}$.