Proof of the optimality of the empirical star algorithm

J.-Y. Audibert

Abstract

This note contains the proof of the assertion made in page 5 of the NIPS paper "Progressive mixture rules are deviation suboptimal". Specifically, it proves that the empirical star algorithm is deviation optimal for the model selection type aggregation problem.

CONTENTS

1. CONTEXT

Let $(\mathcal{X}, \mathcal{B})$ be a measurable space. Let $g_1, \dots, g_d$ be uniformly bounded measurable functions from $\mathcal{X}$ to the set of real numbers $\mathbb{R}$ equipped with its Borel algebra $\mathcal{A}$. Let P be an unknown distribution on $(\mathcal{X} \times \mathbb{R}, \mathcal{B} \otimes \mathcal{A})$ such that the second marginal admits a finite second moment: $\mathbb{E}_{(X,Y)\sim P} Y^2 < \infty$. We observe an i.i.d. sample from P, denoted $Z_1 = (X_1, Y_1), \dots, Z_n = (X_n, Y_n)$. For any measurable function $g : \mathcal{X} \to \mathbb{R}$ such that $\mathbb{E}_{(X,Y)\sim P} g^2(X) < \infty$, we define its risk by

$$R(g) = \mathbb{E}_{(X,Y)\sim P}[Y - g(X)]^2,$$

and its empirical risk by

$$r(g) = \frac{1}{n}\sum_{i=1}^{n}[Y_i - g(X_i)]^2.$$

Introduce

$$g^*_{\mathbf{MS}} \in \operatorname*{argmin}_{g \in \{g_1,\dots,g_d\}} R(g).$$

We want to infer from the sample a function $\hat{g}_{Z_1,\dots,Z_n}$, simply denoted $\hat{g}$ for brevity, such that for any $\varepsilon > 0$, with probability at least $1-\varepsilon$ (with respect to the distribution $P^{\otimes n}$ of the sample), we have

$$R(\hat{g}) - R(g^*_{\mathbf{MS}}) \leq C\frac{\log(d\varepsilon^{-1})}{n}. \tag{1.1}$$

for some constant C, where e is the exponential number.

2. The empirical star algorithm

To achieve (1.1), we propose the following algorithm. Let $\hat{g}^{(\text{erm})}$ be an empirical risk minimizer among the reference functions:

$$\hat{g}^{(\text{erm})} \in \operatorname*{argmin}_{g \in \{g_1,\dots,g_d\}} r(g).$$

For any measurable functions g', g'' from $\mathcal{X}$ to $\mathbb{R}$, let $[g', g'']$ denote the set of functions which are convex combination of g' and g'': $[g', g''] = \{\alpha g' + (1-\alpha)g'' : \alpha \in [0,1]\}$. The empirical star estimator $\hat{g}^{(\text{star})}$ minimizes the empirical risk over a star-shaped set of functions, precisely:

$$\hat{g}^{(\text{star})} \in \operatorname*{argmin}_{g \in [\hat{g}^{(\text{erm})}, g_1] \cup \dots \cup [\hat{g}^{(\text{erm})}, g_d]} r(g).$$

3. The main result

Theorem 1 *Assume that* $|Y| \leq 1$ *almost surely and* $\|g_j\|_\infty \leq 1$ *for any* $j \in \{1,\dots,d\}$. *Then the empirical star algorithm satisfies: for any* $\varepsilon > 0$, *with probability at least* $1-\varepsilon$, *we have*

$$R(\hat{g}^{(\text{star})}) - R(g^*_{\mathbf{MS}}) \leq \frac{200\log[3d(d-1)\varepsilon^{-1}]}{n}.$$

The above inequality naturally implies that Inequality (1.1) holds with $C = 600$.

4. An intermediate result on empirical risk minimization on a segment

Let g' and g'' be two measurable functions from $\mathcal{X}$ to $\mathbb{R}$ that are uniformly bounded by 1: $\|g'\|_\infty \leq 1$ and $\|g''\|_\infty \leq 1$. Let $\hat{g}$ be the empirical risk minimizer on $[g', g'']$, i.e. $\hat{g} \in \underset{g \in [g',g'']}{\operatorname{argmin}} \; r(g)$. Let $\bar{g}$ be the risk minimizer on $[g', g'']$, i.e. $\bar{g} \in \underset{g \in [g',g'']}{\operatorname{argmin}} \; R(g)$.

Theorem 2 *Assume that $|Y| \leq 1$ almost surely. Then for any $\varepsilon > 0$, with probability at least $1 - \varepsilon$, we have simultaneously*

$$R(\hat{g}) - R(\bar{g}) \leq 71 \frac{\log(3\varepsilon^{-1})}{n}. \tag{4.1}$$

and

$$r(\bar{g}) - r(\hat{g}) \leq 54 \frac{\log(3\varepsilon^{-1})}{n}. \tag{4.2}$$

It is likely that the result extends to empirical risk minimization on the convex set of d functions with $d > 2$, but this is out of the scope of this note.

Proof. Let $\hat{t} \in [-1, 1]$ be such that $\hat{g} = \bar{g} + \hat{t}(g'' - g')$. The starting point are the equalities

$$R(\hat{g}) - R(\bar{g}) = 2\hat{t}\mathbb{E}[Y - \bar{g}(X)][g'(X) - g''(X)] + \hat{t}^2\mathbb{E}[g''(X) - g'(X)]^2 \tag{4.3}$$

and

$$r(\hat{g}) - r(\bar{g}) = 2\hat{t}\frac{\sum_{i=1}^n [Y_i - \bar{g}(X_i)][g'(X_i) - g''(X_i)]}{n} + \hat{t}^2 \frac{\sum_{i=1}^n [g''(X_i) - g'(X_i)]^2}{n} \tag{4.4}$$

The following Bernstein's type lemma will be useful [2, Lemma 5].

Lemma 3 *Let $W, W_1, \dots, W_n$ be i.i.d. random variables with $W \leq b$ almost surely and $\mathbb{E}W^2 < \infty$. For any $\varepsilon > 0$, with probability at least $1 - \varepsilon$, we have*

$$\frac{\sum_{i=1}^n W_i}{n} - \mathbb{E}W \leq \sqrt{\frac{2\log(\varepsilon^{-1})\mathbb{E}W^2}{n}} + \max(b, 0)\frac{\log(\varepsilon^{-1})}{3n}.$$

Let $L = \log(3\varepsilon^{-1})/n$ and $\mathcal{D} = \mathbb{E}[g'(X) - g''(X)]^2$. By using the previous lemma to the random variables $[Y_i - \bar{g}(X_i)][g'(X_i) - g''(X_i)]$, $[Y_i - \bar{g}(X_i)][g''(X_i) - g'(X_i)]$, $-[g'(X_i) - g''(X_i)]^2$ and $[g'(X_i) - g''(X_i)]^2$, and using a union bound, for any $\varepsilon > 0$, with probability at least $1 - \varepsilon$, we have simultaneously

$$\frac{\sum_{i=1}^n [Y_i - \bar{g}(X_i)][g'(X_i) - g''(X_i)]}{n}$$

$$- \mathbb{E}[Y-\bar{g}(X)][g'(X)-g''(X)] \le \sqrt{8\mathcal{D}L} + \frac{4L}{3}, \quad (4.5)$$

$$\frac{\sum_{i=1}^n [Y_i-\bar{g}(X_i)][g''(X_i)-g'(X_i)]}{n} - \mathbb{E}[Y-\bar{g}(X)][g''(X)-g'(X)] \le \sqrt{8\mathcal{D}L} + \frac{4L}{3}, \quad (4.6)$$

$$-\frac{\sum_{i=1}^n [g''(X_i)-g'(X_i)]^2}{n} + \mathbb{E}[g''(X)-g'(X)]^2 \le \sqrt{8\mathcal{D}L}. \quad (4.7)$$

In the following, we prove that in the event (of probability at least $1-\varepsilon$) for which these three inequalities hold, Inequalities (4.1) and (4.2) hold. First note that they trivially hold for $\hat{g} = \bar{g}$. We thus focus on the event in which $\hat{g} \neq \bar{g}$ and the above three inequalities hold. In this event, we necessarily have

$$\big(\mathbb{E}[Y-\bar{g}(X)][g'(X)-g''(X)]\big) \frac{\sum_{i=1}^n [Y_i-\bar{g}(X_i)][g'(X_i)-g''(X_i)]}{n} \le 0,$$

hence by using (4.5) and (4.6), we get

$$|\mathbb{E}[Y-\bar{g}(X)][g'(X)-g''(X)]| \le \sqrt{8\mathcal{D}L} + \frac{4L}{3}, \quad (4.8)$$

and

$$\left| \frac{|\sum_{i=1}^n [Y_i-\bar{g}(X_i)][g''(X_i)-g'(X_i)]|}{n} \right| \le \sqrt{8\mathcal{D}L} + \frac{4L}{3}. \quad (4.9)$$

Plugging (4.8) into (4.3), we get

$$R(\hat{g}) - R(\bar{g}) \le 2|\hat{t}| \left(\sqrt{8\mathcal{D}L} + \frac{4L}{3} \right) + \hat{t}^2 \mathcal{D}. \quad (4.10)$$

This allows to get Inequality (4.1) when $\mathcal{D} \le 35L$ since $|\hat{t}| \le 1$. Let us now consider that $\mathcal{D} > 35L$. Then, from (4.7), we have

$$\frac{\sum_{i=1}^n [g''(X_i)-g'(X_i)]^2}{n} \ge \mathcal{D} - \sqrt{8\mathcal{D}L} > 0. \quad (4.11)$$

Besides, by definition of the empirical risk minimizer, (4.4) implies

$$|\hat{t}| \le \frac{|\sum_{i=1}^n [Y_i-\bar{g}(X_i)][g''(X_i)-g'(X_i)]|}{\sum_{i=1}^n [g''(X_i)-g'(X_i)]^2}, \quad (4.12)$$

which, by using (4.9) and (4.11), implies

$$|\hat{t}| \le \frac{\sqrt{8\mathcal{D}L} + \frac{4L}{3}}{\mathcal{D} - \sqrt{8\mathcal{D}L}}.$$

Combining this inequality with (4.10), we obtain

$$R(\hat{g}) - R(\bar{g}) \leq \left(\frac{\sqrt{8\mathcal{D}L} + \frac{4L}{3}}{\mathcal{D} - \sqrt{8\mathcal{D}L}}\right)^2 \left(3\mathcal{D} - 2\sqrt{8\mathcal{D}L}\right) < 71L.$$

Still on the event in which $\hat{g} \neq \bar{g}$ and (4.5), (4.6) and (4.7) hold, let us now prove (4.2). Let $\hat{\mathcal{D}} = \frac{\sum_{i=1}^{n}[g''(X_i) - g'(X_i)]^2}{n}$. First the inequality is trivial for $\hat{\mathcal{D}} = 0$. Now when $\hat{\mathcal{D}} \neq 0$, (4.12) and (4.9) imply

$$|\hat{t}| \leq \frac{\sqrt{8\mathcal{D}L} + \frac{4L}{3}}{\hat{\mathcal{D}}}.$$

Combining (4.9) and (4.4), we have

$$r(\bar{g}) - r(\hat{g}) \leq 2|\hat{t}|\left(\sqrt{8\mathcal{D}L} + \frac{4L}{3}\right) + \hat{t}^2\hat{\mathcal{D}} \leq 3|\hat{t}|\left(\sqrt{8\mathcal{D}L} + \frac{4L}{3}\right).$$

This allows to get Inequality (4.2) when $\mathcal{D} \leq 34.6L$ since $|\hat{t}| \leq 1$. Let us now consider that $\mathcal{D} > 34.6L$. In that case, we again use (4.7) and get

$$r(\bar{g}) - r(\hat{g}) \leq 3\frac{\left(\sqrt{8\mathcal{D}L} + \frac{4L}{3}\right)^2}{\mathcal{D} - \sqrt{8\mathcal{D}L}} \leq 54L,$$

which ends the proof of (4.2). □

5. PROOF OF THE MAIN RESULT

For any j, k in $\{1, \dots, d\}$, introduce

$$\hat{g}_{j,k} \in \underset{g \in [g_j, g_k]}{\text{argmin}}\ r(g)$$

and

$$\bar{g}_{j,k} \in \underset{g \in [g_j, g_k]}{\text{argmin}}\ R(g).$$

Without loss of generality, we may assume that $g^*_{\mathbf{MS}} = g_1$. Let $\hat{e}$ and $\hat{s}$ be such that $\hat{g}^{(\text{erm})} = g_{\hat{e}}$ and $\hat{g}^{(\text{star})} = \hat{g}_{\hat{e},\hat{s}}$.

From Lemma 3 applied to the random variables $[Y_i - \bar{g}_{1,j}(X_i)]^2 - [Y_i - \bar{g}_{j,k}(X_i)]^2$, $j = 1, \dots, d$, $k = 2, \dots, d$ and by using a union bound, for any $\varepsilon_1 > 0$, with probability at least $1 - d(d-1)\varepsilon_1$, for any $j = 1, \dots, d$ and $k = 2, \dots, d$, we simultaneously have

$$R(\bar{g}_{j,k}) - R(\bar{g}_{1,j}) \leq r(\bar{g}_{j,k}) - r(\bar{g}_{1,j}) + \sqrt{\frac{32\mathbb{E}(\bar{g}_{j,k} - \bar{g}_{1,j})^2 \log(\varepsilon_1^{-1})}{n}} + \frac{4\log(\varepsilon_1^{-1})}{3n}.$$

From Theorem 2 and the union bound argument, for any $\varepsilon_2 > 0$, with probability at least $1 - d(d-1)\varepsilon_2/2$, for any $1 \leq j < k \leq d$, we simultaneously have

$$R(\hat{g}_{j,k}) - R(\bar{g}_{j,k}) \leq 71\frac{\log(3\varepsilon_2^{-1})}{n},$$

and

$$r(\bar{g}_{j,k}) - r(\hat{g}_{j,k}) \leq 54\frac{\log(3\varepsilon_2^{-1})}{n}.$$

Let $\varepsilon > 0$. Define $\mathcal{L} = \frac{\log[3d(d-1)\varepsilon^{-1}]}{n}$. Introduce the event on which the following inequalities simultaneously hold for any $j = 1, \dots, d$ and $k = 2, \dots, d$,

$$R(\bar{g}_{j,k}) - R(\bar{g}_{1,j}) \leq r(\bar{g}_{j,k}) - r(\bar{g}_{1,j}) + \sqrt{32\mathbb{E}(\bar{g}_{j,k} - \bar{g}_{1,j})^2\mathcal{L}} + \frac{4\mathcal{L}}{3}, \quad (5.1)$$

$$R(\hat{g}_{j,k}) - R(\bar{g}_{j,k}) \leq 71\mathcal{L}, \quad (5.2)$$

and

$$r(\bar{g}_{j,k}) - r(\hat{g}_{j,k}) \leq 54\mathcal{L}. \quad (5.3)$$

From the previous PAC bounds by taking $\varepsilon_1 = \frac{\varepsilon}{3d(d-1)}$ and $\varepsilon_2 = \frac{4\varepsilon}{3d(d-1)}$, this event holds with probability at least $1 - \varepsilon$. We work hereafter on this high probability event.

Let $\hat{e}, \hat{s}$ in $\{1, \dots, d\}$ such that $g_{\hat{e}} = \hat{g}^{(\text{erm})}$ and $\hat{g}_{\hat{e},\hat{s}} = \hat{g}^{(\text{star})}$. We distinguish two cases.

First case: $R(\bar{g}_{\hat{e},\hat{s}}) \leq R(g_1) + 107\mathcal{L}$.

Then we have

$$R(\hat{g}^{(\text{star})}) = R(\hat{g}_{\hat{e},\hat{s}}) \leq R(\bar{g}_{\hat{e},\hat{s}}) + 71\mathcal{L} \leq R(g_1) + 200\mathcal{L}.$$

Second case: $R(\bar{g}_{\hat{e},\hat{s}}) > R(g_1) + 107\mathcal{L}$.

Introduce $\hat{a} = R(g_{\hat{e}}) - R(\bar{g}_{1,\hat{e}})$ and $\hat{b} = R(g_1) - R(\bar{g}_{1,\hat{e}}) \leq \hat{a}$ (see Figure 1). From (5.2), we have

$$\begin{aligned} R(\hat{g}^{(\text{star})}) - R(g_1) &= R(\hat{g}^{(\text{star})}) - R(\bar{g}_{\hat{e},\hat{s}}) + R(\bar{g}_{\hat{e},\hat{s}}) - R(\bar{g}_{1,\hat{e}}) + R(\bar{g}_{1,\hat{e}}) - R(g_1) \\ &\leq 71\mathcal{L} + \hat{a} - \hat{b} \quad (5.4) \end{aligned}$$

From (5.1), we have

$$\hat{b} < R(\bar{g}_{\hat{e},\hat{s}}) - R(\bar{g}_{1,\hat{e}}) - 107\mathcal{L} \leq r(\bar{g}_{\hat{e},\hat{s}}) - r(\bar{g}_{1,\hat{e}}) + \sqrt{32\mathbb{E}(\bar{g}_{\hat{e},\hat{s}} - \bar{g}_{1,\hat{e}})^2\mathcal{L}} - \frac{317}{3}\mathcal{L},$$

with $r(\bar{g}_{\hat{e},\hat{s}}) - r(\bar{g}_{1,\hat{e}}) \leq r(\bar{g}_{\hat{e},\hat{s}}) - r(\hat{g}^{(\text{star})}) \leq 54\mathcal{L}$ and

$$\begin{aligned} \mathbb{E}(\bar{g}_{\hat{e},\hat{s}} - \bar{g}_{1,\hat{e}})^2 &\leq 2\mathbb{E}(\bar{g}_{\hat{e},\hat{s}} - g_{\hat{e}})^2 + 2\mathbb{E}(g_{\hat{e}} - \bar{g}_{1,\hat{e}})^2 \\ &\leq 2[R(g_{\hat{e}}) - R(\bar{g}_{\hat{e},\hat{s}})] + 2\hat{a} \end{aligned}$$

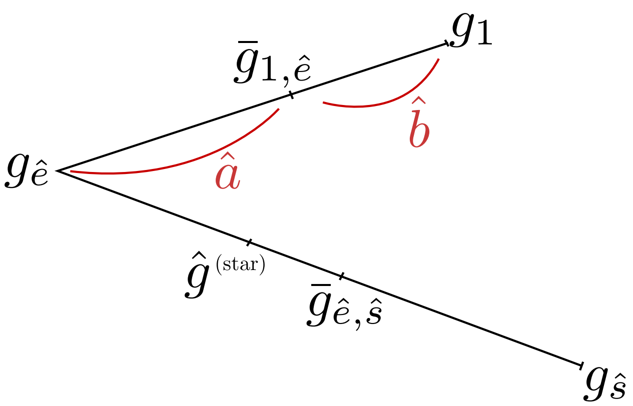

Figure 1: the second case configuration

$$
\begin{aligned}
&\leq 2[R(g_{\hat{e}}) - R(g_1) - 107\mathcal{L}] + 2\hat{a} \\
&= 4\hat{a} - 2\hat{b} - 214\mathcal{L}.
\end{aligned}
$$

Consequently, we have

$$
\hat{b} < R(\bar{g}_{\hat{e},\hat{s}}) - R(\bar{g}_{1,\hat{e}}) \leq -\frac{155}{3}\mathcal{L} + \sqrt{128\hat{a}\mathcal{L} - 64\hat{b}\mathcal{L} - 6848\mathcal{L}^2}. \quad (5.5)
$$

From (5.1), noting that $\bar{g}_{\hat{e},\hat{e}} = g_{\hat{e}}$, we also have

$$
\begin{aligned}
\hat{a} = R(g_{\hat{e}}) - R(\bar{g}_{1,\hat{e}}) &\leq r(g_{\hat{e}}) - r(\bar{g}_{1,\hat{e}}) + \sqrt{32\mathbb{E}(g_{\hat{e}} - \bar{g}_{1,\hat{e}})^2\mathcal{L}} + \frac{4\mathcal{L}}{3} \\
&\leq r(g_{\hat{e}}) - r(\bar{g}_{1,\hat{e}}) + \sqrt{32\hat{a}\mathcal{L}} + \frac{4\mathcal{L}}{3},
\end{aligned}
$$

and

$$
\begin{aligned}
r(g_{\hat{e}}) - r(\bar{g}_{1,\hat{e}}) &\leq r(g_1) - r(\bar{g}_{1,\hat{e}}) \\
&\leq R(g_1) - R(\bar{g}_{1,\hat{e}}) + \sqrt{32\mathbb{E}(g_1 - \bar{g}_{1,\hat{e}})^2\mathcal{L}} + \frac{4\mathcal{L}}{3} \\
&\leq \hat{b} + \sqrt{32\hat{b}\mathcal{L}} + \frac{4\mathcal{L}}{3}
\end{aligned}
$$

Combining the last two inequalities, we get

$$
\hat{a} \leq \hat{b} + \sqrt{32\hat{b}\mathcal{L}} + \sqrt{32\hat{a}\mathcal{L}} + \frac{8\mathcal{L}}{3},
$$

hence $\hat{a} \leq 16 + v + 2\sqrt{64 + 8v}$, with $v = \hat{b} + \sqrt{32\hat{b}\mathcal{L}} + \frac{8\mathcal{L}}{3}$. Plugging this inequality into (5.5) and solving the inequation gives $\hat{b} \leq 64\mathcal{L}$, hence $\hat{a} - \hat{b} \leq 129\mathcal{L}$. From (5.4), this gives the desired result.

6. Extensions

Consider a (loss) function $\ell : [-1,1] \times [-1,1] \to \mathbb{R}$ such that there exist $b_1 > 0$ and $b_2 > 0$ for which for any $y \in [-1,1]$, the function $\ell_y : y' \mapsto \ell(y,y')$ is twice differentiable and satisties: for any $y' \in [-1,1]$, $\ell_{y'}(y') = 0$ and

$$b_1 \leq \ell''_y(y') \leq b_2.$$

The results can be extended to the setting where the least square risk is replaced by

$$R(g) = \mathbb{E}_{(X,Y)\sim P}\ell[Y, g(X)].$$

Besides, for the least square risk, we have considered boundedness assumptions, which can be weakened. In particular, the results still hold (up to modification of the constant factors) if we assume that for some positive real numbers a, M, B,

$$\mathbb{E}\big(e^{a|Y-\mathbb{E}(Y|X=x)|}\big|X=x\big) \leq M \qquad \text{for any } x \in \mathcal{X}$$

and

$$\|g' - g''\|_\infty \leq B \qquad \text{for any } g', g'' \text{ in } \{g_1, \dots, g_d\} \cup \{x \mapsto \mathbb{E}(Y|X=x)\}.$$

To get this extension, one can use a similar analysis to the one done in Section 7.1.1 of [1, Chapter 1].

References

[1] J.-Y. Audibert. *PAC-Bayesian statistical learning theory*. PhD thesis, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7, 2004. `http://certis.enpc.fr/~audibert/ThesePack.zip`.

[2] J.-Y. Audibert, R. Munos, and Cs. Szepesvári. Exploration-exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410:1876–1902, 2009.