Occam's Hammer

Gilles Blanchard[1] and François Fleuret[2]

[1] Fraunhofer FIRST.IDA, Berlin, Germany blanchar@first.fraunhofer.de
[2] EPFL, CVLAB, Lausanne, Switzerland francois.fleuret@epfl.ch

Abstract. We establish a generic theoretical tool to construct probabilistic bounds for algorithms where the output is a subset of objects from an initial pool of candidates (or more generally, a probability distribution on said pool). This general device, dubbed "Occam's hammer", acts as a meta layer when a probabilistic bound is already known on the objects of the pool taken individually, and aims at controlling the *proportion* of the objects in the set output not satisfying their individual bound. In this regard, it can be seen as a non-trivial generalization of the "union bound with a prior" ("Occam's razor"), a familiar tool in learning theory. We give applications of this principle to randomized classifiers (providing an interesting alternative approach to PAC-Bayes bounds) and multiple testing (where it allows to retrieve exactly and extend the so-called Benjamini-Yekutieli testing procedure).

1 Introduction

In this paper, we establish a generic theoretical tool allowing to construct probabilistic bounds for algorithms which take as input some (random) data and return as an output a set A of objects among a pool $\mathcal{H}$ of candidates (instead of a single object $h \in \mathcal{H}$ in the classical setting). Here the "objects" could be for example classifiers, functions, hypotheses... according to the setting. One wishes to predict that each object h in the output set A satisfies a property $R(h, \alpha)$ (where α is an ajustable level parameter); the purpose of the probabilistic bound is to guarantee that the proportion of objects in A for which the prediction is false does not exceed a certain value, and this with a prescribed statistical confidence $1 - \delta$. Our setting also covers the more general case where the algorithm returns a (data-dependent) probability density over $\mathcal{H}$.

Such a wide scope can appear dubious in its generality at first and even seem to border with abstract nonsense, so let us try to explain right away what is the nature of our result, and pinpoint a particular example to fix ideas. The reason we encompass such a general framework is that our result acts as a 'meta' layer: we will pose that we already have at hand a probabilistic bound for single, fixed elements $h \in \mathcal{H}$. Assuming the reader is acquainted with classical learning theory, let us consider the familiar example where $\mathcal{H}$ is a set of classifiers and we observe an i.i.d. labeled sample of training data as an input. For each fixed classifier $h \in \mathcal{H}$, we can predict with success probability at least $1-\delta$ the property $R(h, \delta)$ that the generalization error of h is bounded by the training error up to

a quantity $\varepsilon(\delta)$, for example using the Chernoff bound. In the classical setting, a learning method will return a single classifier $h \in \mathcal{H}$. If nothing is known about the algorithm, we have to resort to worst-case analysis, that is, obtain a uniform bound over $\mathcal{H}$; or in other terms, ensure that the probability that the predicted properties hold for *all* $h \in \mathcal{H}$ is at least $1 - \delta$. The simplest way to achieve this is to apply the union bound, combined with a prior Π on $\mathcal{H}$ (assumed to be countable in this situation) prescribing how to distribute the failure probability δ over $\mathcal{H}$. In the folklore, this is generally referred to as *Occam's razor* bound, because the quantity $-\log(\Pi(h))$, which can be interpreted as a coding length for objects $h \in \mathcal{H}$, appears in some explicit forms of the bound. This can be traced back to [4] where the motivations and framework were somewhat different. The formulation we use here seems to have first appeared explicitly in [9] .

The goal of the present work is to put forward what can be seen as an analogue of the above "union bound with a prior" for the set output (or probability output) case, which we call *Occam's hammer* by remote analogy with the principle underlying Occam's razor bound. Occam's hammer relies on *two* priors: a complexity prior similar to the razor's (except it can be continuous) and a second prior over the output set size or inverse output density. We believe that Occam's hammer is not as immediately straightforward as the classical union bound, and hope to show that it has potential for interesting applications. For reasons of space, we will cut to the chase and first present Occam's hammer in an abstract setting in the next section (the reader should keep in mind the classifiers example to have a concrete instance at hand) then proceed to some applications in Section 3 (including a detailed treatment of the classifiers example in Section 3.1) and a discussion about tightness in Section 4. A natural application field is *multiple testing*, where we want to accept or reject (in the classical statistical sense) hypotheses from a pool $\mathcal{H}$; this will be developed in section 3.2. The present work was motivated by the PASCAL theoretical challenge [6] on this topic.

2 Main result

2.1 Setting

Assume we have a pool of objects which is a measurable space $(\mathcal{H}, \mathfrak{H})$ and observe a random variable X (which can possibly represent an entire data sample) from a probability space $(\mathcal{X}, \mathfrak{X}, P)$. Our basic assumption is:

Assumption A: for every $h \in \mathcal{H}$, and $\delta \in [0, 1]$, we have at hand a set $\mathcal{B}(h, \delta) \in \mathfrak{X}$ such that $\mathbb{P}_{X \sim P}[X \in \mathcal{B}(h, \delta)] \leq \delta$. We call $\mathcal{B}(h, \delta)$ "bad event at level δ for h". Moreover, we assume that the function $(x, h, \delta) \in \mathcal{X} \times \mathcal{H} \times [0, 1] \mapsto \mathbf{1}\{x \in \mathcal{B}(h, \delta)\}$ is jointly measurable in its three variables (this amounts to say that the set defined by this indicator function is measurable in the product space). Finally, we assume that for any $h \in \mathcal{H}$ we have $\mathcal{B}(h, 0) = \emptyset$.

It should be understood that "bad events" represent regions where a certain desired property does not hold, such as the true error being larger than the

empirical error plus $\varepsilon(\delta)$ in the classification case. Note that this 'desirable property' implicitly depends on the assigned confidence level $1-\delta$. We should keep in mind that as δ decreases, the set of observations satisfying the corresponding property grows larger, but the property itself loses significance (as is clear once again in the generalization error bound example). Of course, the 'properties' corresponding to $\delta = 0$ or 1 will generally be trivial ones, i.e. $\mathcal{B}(h,0) \equiv \emptyset$ and $\mathcal{B}(h,1) \equiv \mathcal{X}$. Let us reformulate the union bound in this setting:

Proposition 2.1 (Abstract Occam's razor). *Let Π be a prior probability distribution on $\mathcal{H}$ and assume* **(A)** *holds. Then*

$$\mathbb{P}_{X\sim P}\left[\exists h \in \mathcal{H}, X \in \mathcal{B}(h, \delta\Pi(\{h\}))\right] \leq \delta. \tag{1}$$

The following formulation is equivalent: for any rule taking X as an input and returning $h_X \in \mathcal{H}$ as an output (in a measurable way as a function of X), we have

$$\mathbb{P}_{X\sim P}\left[X \in \mathcal{B}(h_X, \delta\Pi(\{h_X\}))\right] \leq \delta. \tag{2}$$

Proof. In the first inequality we want to bound the probability of the event

$$\bigcup_{h\in\mathcal{H}} \mathcal{B}(h, \delta\Pi(\{h\}))\,.$$

Since we assumed $B(h,0) = \emptyset$ the above union can be reduced to a countable union over the set $\{h \in \mathcal{H} : \Pi(\{h\}) > 0\}$. It is in particular measurable. Then, we apply the union bound over the sets in this union. The event in the second inequality can be written as

$$\bigcup_{h\in\mathcal{H}} \left(\{X : h_X = h\} \cap \mathcal{B}(h, \delta\Pi(\{h\}))\right).$$

It is measurable by the same argument as above, and a subset of the first considered event. Finally, from the second inequality we can recover the first one by considering a "rule" that for any X returns an element of $\{h \in \mathcal{H} | X \in \mathcal{B}(h, \delta\Pi(\{h\}))\}$ if this set is non empty, and some arbitrary fixed h_0 otherwise. It is possible to do so in a measurable way again because the set of atoms of Π is countable. □

Note that Occam's razor is obviously only interesting for *atomic* priors, and therefore essentially only useful for a countable object space $\mathcal{H}$.

2.2 False prediction rate

Let us now assume that we have an algorithm or "rule" taking X as an input and returning as an output a subset $A_X \subset \mathcal{H}$; we assume the function $(X,h) \in \mathcal{X}\times\mathcal{H} \mapsto \mathbf{1}\{h \in A_X\}$ is jointly measurable in its two variables. What we are interested in is upper bounding the proportion of objects in A_X falling in a "bad event". Here the word 'proportion' refers to a volume ratio, where volumes are measured through a reference measure Λ on $(\mathcal{H}, \mathfrak{H})$. Like in Occam's razor, we want to allow the set level to depend on h and possibly on A_X. Here is a formal definition for this:

Definition 2.2 (False prediction rate, FPR) *Pose assumption* **(A)**. *Fix a function* $\Delta : \mathcal{H} \times \mathbb{R}_+ \to [0,1]$*, jointly measurable in its two parameters, called the* level function*. Let* Λ *be a volume measure on* $\mathcal{H}$*; we adopt the notation* $|S| \equiv \Lambda(S)$ *for* $S \in \mathfrak{H}$*. Define the* false prediction rate *for level function* Δ *as*

$$\rho_\Delta(X, A) = \frac{|A \cap \{h \in \mathcal{H} : X \in \mathcal{B}(h, \Delta(h, |A|))\}|}{|A|}, \text{ if } |A| \in (0, \infty); \qquad (3)$$

and $\rho_\Delta(X, A) = 0$*, if* $|A| = 0$ *or* $|A| = \infty$.

The name *false prediction rate* was chosen by reference to the notion of *false discovery rate* (FDR) for multiple testing (see below Section 3.2). We will drop the index Δ to lighten notation when it is unambiguous. The pointwise false prediction rate for a specific algorithm $X \mapsto A_X$ is therefore $\rho(X, A_X)$. In what follows, we will actually upper bound the *expected value* $\mathbb{E}_X[\rho(X, A_X)]$ over the drawing of X. In some cases, controlling the averaged FPR is a goal of its own right. Furthermore, if we have a bound on $\mathbb{E}_X[\rho]$, then we can apply straightforwardly Markov's inequality to obtain a confidence bound over ρ:

$$\mathbb{E}_X[\rho(X, A_X)] \leq \gamma \Rightarrow \rho(X, A_X) \leq \gamma\delta^{-1} \text{ with probability } 1 - \delta.$$

2.3 Warming up: algorithm with constant volume output

To begin with, let us consider the easier case where the set ouput given by the algorithm has a fixed size, i.e. $|A_X| = a$ is a constant instead of being random.

Proposition 2.3. *Suppose assumption* **(A)** *holds and that* $(X, h) \in \mathcal{X} \times \mathcal{H} \mapsto \mathbf{1}\{h \in A_X\}$ *is jointly measurable in its two variables. Assume* $|A_X| = \Lambda(A_X) \equiv a$ *a.s. Let* π *be a probability density function on* $\mathcal{H}$ *with respect to the measure* Λ*. Then putting* $\Delta(h, |A|) = \min(\delta a \pi(h), 1)$*, it holds that*

$$\mathbb{E}_{X \sim P}[\rho(X, A_X)] \leq \delta.$$

Proof: Obviously, Δ is jointly measurable in its two variables, and by the composition rule so is the function $X \mapsto \rho(X, A_X)$. We then have

$$\begin{aligned}
\mathbb{E}_{X \sim P}[\rho(X, A_X)] &= \mathbb{E}_{X \sim P}\left[a^{-1} |A_X \cap \{h \in \mathcal{H}, X \in \mathcal{B}(h, \Delta(h, |A_X|))\}|\right] \\
&\leq \mathbb{E}_{X \sim P}\left[|\{h \in \mathcal{H} : X \in \mathcal{B}(h, \min(\delta a \pi(h), 1))\}|\right] a^{-1} \\
&= \mathbb{E}_{X \sim P}\left[\int_h \mathbf{1}\{X \in \mathcal{B}(h, \min(\delta a \pi(h), 1))\} d\Lambda(h)\right] a^{-1} \\
&= \int_h \mathbb{P}_{X \sim P}[X \in \mathcal{B}(h, \min(\delta a \pi(h), 1))] \, d\Lambda(h) a^{-1} \\
&\leq \delta \int_h \pi(h) d\Lambda(h) = \delta. \qquad \square
\end{aligned}$$

As a sanity check, consider a countable set $\mathcal{H}$ with Λ the counting measure, and an algorithm returning only singletons, $A_X = \{h_X\}$, so that $|A_X| \equiv 1$. Then in this case $\rho \in \{0, 1\}$, and with the above choice of Δ, we get $\rho(X, \{h\}) = \mathbf{1}\{X \in \mathcal{B}(h, \delta\pi(h))\}$. Therefore, $\mathbb{E}_X[\rho(X, A_X)] = \mathbb{P}_X[X \in \mathcal{B}(h_X, \delta\pi(h_X))] \leq \delta$, i.e., we have recovered version (2) of Occam's razor.

2.4 General case

The previous section might let us hope that $\Delta(h, |A|) = \delta|A|\pi(h)$ would be a suitable level function in the more general situation where the size $|A_X|$ is also variable; but things get more involved. The observant reader might have noticed that, in Proposition 2.3, the weaker assumption $|A_X| \geq a$ a.s. is actually sufficient to ensure the conclusion. This thefore suggests the following strategy to deal with variable size of A_X: (1) consider a discretization of sizes through a decreasing sequence (a_k) converging to zero; and a prior Γ on the elements of the sequence; (2) apply Proposition 2.3 for all k with $(a_k, \Gamma(a_k)\delta)$ in place of (a, δ); (3) define $\Delta(h, |A|) = \delta\pi(h)a_k\Gamma(a_k)$ whenever $|A| \in [a_k, a_{k-1})$; then by summation over k (or, to put it differently, the union bound) it holds that $\mathbb{E}[\rho] \leq \delta$ for this choice of Δ.

This is a valid approach, but we will not enter into more details concerning it; rather, we propose what we consider to be an improved and more elegant result below, which will additionally allow to handle the more general case where the algorithm returns a probability density over $\mathcal{H}$ instead of just a subset. However, we will require a slight strengthening of assumption **(A)**:

Assumption A': like assumption **(A)**, but we additionaly require that for any $h \in \mathcal{H}$, $\mathcal{B}(h, \delta)$ is a nondecreasing sequence of sets as a function of δ, i.e., $\mathcal{B}(h, \delta) \subset \mathcal{B}(h, \delta')$ for $\delta \leq \delta'$.

The assumption of nondecreasing bad events as a function of their probability seems quite natural and is satisfied in the applications we have in mind; in classification for example, bounds on the true error are nonincreasing in the parameter δ (so the set of samples where the bound is violated is nondecreasing). We now state our main result (proof found in the appendix):

Theorem 2.4 (Occam's hammer). *Pose assumption* **(A')** *satisfied. Let:*

(i) Λ be a nonnegative reference measure on $\mathcal{H}$ (the volumic measure*);*

(ii) Π be a probability distribution on $\mathcal{H}$ absolutely continuous wrt Λ (the complexity prior*), and denote $\pi = \frac{d\Pi}{d\Lambda}$;*

(iii) Γ be a probability distribution on $(0, +\infty)$ (the inverse density prior*). Put $\beta(x) = \int_0^x u d\Gamma(u)$ for $x \in (0, +\infty)$. Define the level function*

$$\Delta(h, \theta) = \min(\delta\pi(h)\beta(\theta^{-1}), 1).$$

Then for any algorithm $X \mapsto \theta_X$ returning a probability density θ_X over $\mathcal{H}$ with respect to Λ, and such that $(X, h) \mapsto \theta_X(h)$ is jointly measurable in its two variables, it holds that

$$\mathbb{P}_{X\sim P, h\sim\Theta_X}\left[X \in \mathcal{B}(h, \Delta(h, \theta_X(h)))\right] \leq \delta\,,$$

where Θ_X is the distribution on $\mathcal{H}$ such that $\frac{d\Theta_X}{d\Lambda} = \theta_X$.

Comments: The conclusion of the above theorem is a probabilistic statement over the *joint* draw of the input variable X and the object h, where the conditional distribution of h given X is Θ_X .

Note that a rule returning a probability density distribution over $\mathcal{H}$ is more general than a rule returning a set, as the latter case can be cast into the former by considering a constant density over the set, $\theta_A(h) = |A|^{-1}\mathbf{1}\{h \in A\}$; in this case the inner probability over $h \sim \Theta_{A_X}$ is exactly the false prediction rate $\rho_\Delta(X, A_X)$ introduced previously. This specialization gives a maybe more intuitive interpretation of the inverse density prior Γ, which then actually becomes a prior on the volume of the set output. We can thus recover the case of constant set volume a of Proposition 2.3 by using the above specialization and taking a Dirac distribution for the inverse density prior, $\Gamma = \delta_a$. In particular, version (2) of Occam's razor is a specialization of Occam's hammer (up to the minor strengthening in assumption **(A')**).

To compare with the "naive" strategy described earlier based on a size discretization sequence (a_k), we get the following advantages: Occam's hammer also works with the more general case of a probability output; it avoids any discretization of the prior; finally, if even we take the discrete prior $\Gamma = \sum_k \gamma_k \delta_{a_k}$ in Occam's hammer, the level function for $|A| \in [a_k, a_{k-1})$ will be proportional to the partial sum $\sum_{j \leq k} \gamma_j a_j$, instead of only the term $\gamma_k a_k$ in the naive approach (remember that the higher the level function, the better, since the corresponding 'desirable property' is more significant for higher levels).

3 Applications

3.1 Randomized classifiers: an alternate look at PAC-Bayes bounds

Our first application is concerned with our running example, classifiers. More precisely, assume the input variable is actually an i.i.d. sample $S = (X_i, Y_i)_{i=1}^n$, and $\mathcal{H}$ is a set of classifiers. Let $\mathcal{E}(h)$, resp. $\widehat{\mathcal{E}}(h, S)$, denote the generalization, resp. training, error. We assume that generalization and training error are measurable in their respective variables, which is a tame assumption for all practical purposes. We consider a randomized classification algorithm, consisting in selecting a probability density function θ_S on $\mathcal{H}$ based on the sample (again, jointly measurable in (x, h)), then drawing a classifier at random from $\mathcal{H}$ using the distribution Θ_S such that $\frac{d\Theta_S}{d\Lambda} = \theta_s$, where Λ is here assumed to be a reference *probability* measure. For example, we could return the uniform density on the set of classifiers $A_S \subset \mathcal{H}$ having their empirical error less than a (possibly data-dependent) threshold. Combining Occam's Hammer with the Chernoff bound, we obtain the following result:

Proposition 3.1. *Let Λ be a probability measure over $\mathcal{H}$; consider an algorithm $S \mapsto \theta_S$ returning a probability density θ_S over $\mathcal{H}$ (wrt. Λ). Let $\delta \in (0, 1)$ and $k > 0$ be fixed. If h_S is a randomized classifier drawn according to Θ_S, the following inequality holds with probability $1 - \delta$ over the joint draw of S and h_S:*

$$D_+\left(\widehat{\mathcal{E}}(h_S, S) \| \mathcal{E}(h_S)\right) \leq \frac{1}{n}\left(\log\left((k+1)\delta^{-1}\right) + \left(1 + \frac{1}{k}\right)\log_+ \theta_S(h_S)\right), \quad (4)$$

where $\log_+$ is the positive part of the logarithm; and $D_+(q||p) = q \log \frac{q}{p} + (1 - q) \log \frac{1-q}{1-p}$ if $q < p$ and 0 otherwise.

Proof. Define the bad events $\mathcal{B}(h,\delta) = \left\{S : D_+(\widehat{\mathcal{E}}(h,S)\|\mathcal{E}(h)) \leq \frac{\log \delta^{-1}}{n}\right\}$, satisfying assumption **(A')** by Chernoff's bound (see, e.g., [7]), including the measurability assumptions of **(A)** by the composition rule. Choose $\Pi = \Lambda$, i.e., $\pi \equiv 1$, and Γ the probability distribution on $[0,1]$ having density $\frac{1}{k}x^{-1+\frac{1}{k}}$, so that $\beta(x) = \frac{1}{k+1}\min(x^{1+\frac{1}{k}}, 1)$, and apply Occam's hammer. Replacing δ by the level function given by Occam's hammer gives rise to the following factor:

$$\begin{aligned}
\log(\min(\delta\pi(h_S)\beta(\theta_S(h_S)^{-1}), 1)^{-1}) &= \log_+(\delta^{-1}\min((k+1)^{-1}\theta_S(h_S)^{-\frac{k+1}{k}}, 1)^{-1}) \\
&= \log_+(\delta^{-1}\max((k+1)\theta_S(h_S)^{\frac{k+1}{k}}, 1)) \\
&\leq \log_+((k+1)\delta^{-1}\max(\theta_S(h_S)^{\frac{k+1}{k}}, 1)) \\
&\leq \log((k+1)\delta^{-1}) + \log_+(\theta_S(h_S)^{\frac{k+1}{k}}) \\
&= \log((k+1)\delta^{-1}) + \left(1+\frac{1}{k}\right)\log_+(\theta_S(h_S))\,.
\end{aligned}$$

□

Comparison with PAC-Bayes bounds. The by now quite well-established PAC-Bayes bounds ([9], see also [7] and references therein, and [5, 1] for recent developments) deal with a similar setting of randomized classifiers. One important difference is that PAC-Bayes bounds are generally concerned with bounding the *averaged error* $\mathbb{E}_{h\sim\Theta_S}[\mathcal{E}(h)]$ of the randomized procedure. Occam's hammer, on the other hand, bounds directly the true error of a single randomized output: this is particularly relevant in practice since the information given to the user by Occam's hammer bound concerns precisely the classifier returned by the rule. In other words, Proposition 3.1 appears as a *pointwise* version of the PAC-Bayes bound. It is important to understand that a pointwise version is a stronger statement, as we can recover a traditional PAC-Bayes bound as a consequence of Proposition 3.1 (the proof is found in the appendix):

Corollary 3.2. *Provided the conclusion of Proposition 3.1 holds, for any $k > 0$ the following holds with probability δ over the the draw of S :*

$$\begin{aligned}
&D_+\left(\mathbb{E}_{h_S\sim\Theta_S}\left[\widehat{\mathcal{E}}(h_S,S)\right] \middle\| \mathbb{E}_{h_S\sim\Theta_S}\left[\mathcal{E}(h_S)\right]\right) \\
&\qquad\qquad \leq \frac{1}{n}\left(\log\left((k+1)\delta^{-1}\right) + \frac{k+1}{k}KL(\Theta_S\|\Lambda) + 3.5 + \frac{1}{2k}\right),
\end{aligned}$$

where KL denotes the Kullback-Leibler divergence.

It is interesting to compare this to an existing version of the PAC-Bayes bound: if we pick $k = n-1$ in the above corollary, then we recover almost exactly a tight version of the PAC-Bayes bound given in [7], Theorem 5.1 (the differences are: a $(n-1)^{-1}$ instead of n^{-1} factor in front of the KL divergence term, and the additional trailing terms bounded by $\frac{4}{n}$). Hence, Proposition 3.1 proves a stronger property than the latter cited PAC-Bayes bound (admittedly up to the very minor loosening just mentioned).

Note that pointwise results for randomized procedures using the PAC-Bayes approach have already appeared in recent work [1, 5], using a Bernstein type bound rather than Chernoff. It is not clear to us however whether the methodology developed there is precise enough to obtain a Chernoff type bound and recover a pointwise version of [7], Theorem 5.1, which is what we do here.

At any rate, we believe the Occam's hammer approach should turn out more precise for pointwise results. To give some support to this claim, we note that all existing PAC-Bayes bounds up to now structurally rely on Chernoff's method (i.e. using the Laplace transform) via two main ingredients: (1) the entropy extremal inequality $\mathbb{E}_P[X] \geq \log \mathbb{E}_Q\left[e^X\right] + D(P||Q)$ and (2) inequalities on the Laplace transform of i.i.d. sums. Occam's hammer is, in a sense, less sophisticated since it only relies on simple set measure manipulations and contains no intrinsic exponential moment inequality argument. On the other hand, it acts as a 'meta' layer into which any other bound family can be plugged in. These could be bounds based on the Laplace transform (Chernoff method) as above, or not: in the above example, we have used Chernoff's bound for the sake of comparison with earlier work, but we could as well have plugged in the tighter binomial tail inversion bound (which is the most accurate deterministic bound possible for estimating a Bernoulli parameter), and this is clearly a potential improvement for finite size training sets. To this regard, we plan to make an extensive comparison on simulations in future work. In classical PAC-Bayes, there is no such clear separation between the bound and the randomization; they are intertwined in the analysis.

3.2 Multiple testing: a family of "step-up" algorithms with distribution-free FDR control

We now change gears and switch to the context of multiple testing. $\mathcal{H}$ is now a set of *null hypotheses* concerning the distribution P. In this section we will assume for simplicity that $\mathcal{H}$ is finite and the volume measure Λ is the counting measure, although this could be obviously extended. The goal is, based on oberved data, to discover a subset of hypotheses which are predicted to be *false* (or "*rejected*"). To have an example in mind, think of microarray data, where we observe a small number of i.i.d. repetitions of a variable in very high dimension d (the total number of genes), corresponding to the expression level of said genes, and we want to find a set of genes having average expression level bigger than some fixed threshold t. In this case, there is one null hypothesis h per gene, namely that the average expression level for this gene is *lower* than t.

We assume that we already have at hand a family of tests $T(X, h, \alpha)$ of level α for each individual h. That is, $T(X, h, \alpha)$ is a measurable function taking values in $\{0, 1\}$ (the value 1 corresponds to "null hypothesis rejected") such that for all $h \in \mathcal{H}$, for all distributions P such that h is true, $\mathbb{P}_{X\sim P}[T(X, h, \alpha) = 1] \leq \alpha$. To apply Occam's hammer, we suppose that the family $T(X, h, \alpha)$ is increasing, i.e. $\alpha \geq \alpha' \Rightarrow T(X, h, \alpha) \geq T(X, h, \alpha')$. This is generally statisfied, as typically tests have the form $T(X, h, \alpha) = \mathbf{1}\{F(h, X) > \phi(\alpha)\}$, where F is some test statistic

and $\phi(\alpha)$ is a nonincreasing threshold function (as, for example, in a one-sided T-test).

For a fixed, but unknown, data distribution P, let us define

$$\mathcal{H}_0 = \{h \in \mathcal{H} : P \text{ satisfies hypothesis } h\}$$

the set of true null hypotheses, and $\mathcal{H}_1 = \mathcal{H} \setminus \mathcal{H}_0$ its complementary. An important and relatively recent concept in multiple testing is that of *false discovery rate* (FDR) introduced in [2]. Let $A : X \mapsto A_X \subset \mathcal{H}$ be a rule returning a set of rejected hypotheses based on the data. The FDR of such a procedure is defined as

$$FDR(A) = \mathbb{E}_{X \sim P}\left[\frac{|A_X \cap \mathcal{H}_0|}{|A_X|}\right] . \tag{5}$$

Note that, in contrast to our notion of FPR introduced in Section 2.2, the FDR is already an averaged quantity. A desirable goal is to design testing procedures where it can be ensured that the FDR is controlled by some fixed level α. The rationale behind this is that, in practice, one can afford that a small proportion of rejected hypotheses are actually true. Before this notion was introduced, in most cases one would instead bound the probability that *at least one* hypothesis was falsely rejected: this is typically achieved using the (uniform) union bound, known as "Bonferroni's correction" in the multitesting literature. The hope is that, by allowing a little more slack in the acceptable error by controlling only the FDR, one obtains less conservative testing procedures as a counterpart. We refer the reader to [2] for a more extended discussion on these issues.

Let us now describe how Occam's hammer can be put to use here. Let Π be a probability distribution over $\mathcal{H}$, Γ be a probability distribution over the integer inteval $[1 \dots |\mathcal{H}|]$, and $\beta(k) = \sum_{i \leq k} i\Gamma(i)$. Define the procedure returning the following set of hypotheses :

$$A : X \mapsto A_X = \bigcup \{G \subset \mathcal{H} : \ \forall h \in G, \ T(X, h, \alpha\Pi(h)\beta(|G|)) = 1\} . \tag{6}$$

This type of procedure is called "step-up" and can be implemented through a simple water-emptying type algorithm. Namely, it is easy to see that if we define

$$B_\gamma = \{h : T(X, h, \alpha\Pi(h)\gamma) = 1\} \ , \text{ and } \gamma(X) = \sup\{\gamma \geq 0 : \beta(|B_\gamma|) \geq \gamma\} \ ,$$

then $A_X = B_{\gamma(X)}$. The easiest way to construct this is to sort the hypotheses $h \in \mathcal{H}$ by increasing order of their "weighted p-values"

$$p(h, X) = \Pi(h) \inf\{\gamma \geq 0 : T(X, h, \gamma) = 1\} \ ,$$

and to return the $k(X)$ first hypotheses for this order, where $k(X)$ is the largest integer such that $p^{(k)}(X) \leq \alpha\beta(k)$ (where $p^{(k)}(X)$ is the k-th ordered p-value as defined above).

We have the following property for this procedure:

Proposition 3.3. *The set of hypotheses returned by the procedure defined by* (6) *has its false discovery rate bounded by* $\Pi(\mathcal{H}_0)\alpha \leq \alpha$.

Proof. It can be checked easily that $(x,h) \mapsto |A_X|^{-1}\mathbf{1}\{h \in A_X\}$ is measurable in its two variables (this is greatly simplified by the fact that $\mathcal{H}$ is assumed to be finite here). Define the collection of "bad events" $B(h,\delta) = \{X : T(X,h,\delta) = 1\}$ if $h \in \mathcal{H}_0$, and $B(h,\delta) = \emptyset$ otherwise. It is an increasing family by the assumption on the test family. Obviously, for any $G \subset \mathcal{H}$, and any level function Δ:

$$G \cap \{h \in \mathcal{H} : X \in \mathcal{B}(h, \Delta(h,|G|))\} = G \cap \mathcal{H}_0 \cap \{h \in \mathcal{H} : T(X,h,\Delta(h,|G|)) = 1\} \ ;$$

therefore, for any G satifying

$$G \subset \{h \in \mathcal{H} : T(X,h,\Delta(h,|G|)) = 1\} \ , \tag{7}$$

it holds that $|G \cap \{h \in \mathcal{H} : X \in \mathcal{B}(h, \Delta(h,|G|))\}| = |G \cap \mathcal{H}_0|$, so that the averaged (over the draw of X) FPR (3) for level function Δ coincides with the FDR (5). When Δ is nondecreasing in its second parameter, it is straightforward that the union of two sets satisfying (7) also satisfy (7), hence A_X satisfies the above condition for the level function given by Occam's Hammer. Define the modified prior $\widetilde{\Pi}(h) = \mathbf{1}\{h \in \mathcal{H}_0\}\Pi(\mathcal{H}_0)^{-1}\Pi(h)$. Apply Occam's hammer with priors Λ, $\widetilde{\Pi}$, Γ and $\delta = \Pi(\mathcal{H}_0)\alpha$ to conclude. □

Interestingly, the above result specialized to the case where Π is uniform on $\mathcal{H}$ and $\Gamma(i) = \kappa^{-1}i^{-1}$, $\kappa = \sum_{i \leq |\mathcal{H}|} i^{-1}$ results in $\beta(i) = \kappa^{-1}i$, and yields exactly what is known as the *Benjamini-Yekutieli (BY) step-up procedure* [3]. Unfortunately, the interest of the BY procedure is mainly theoretical, because the more popular *Benjamini-Hochberg* (BH) step-up procedure [2] is generally preferred in practice. The BH procedure is in all points similar to BY, except the above constant κ is replaced by 1. The BH procedure was shown to result in controlled FDR at level α *if the test statistics are independent or satisfy a certain form of positive dependency* [3]. In contrast, the BY procedure is distribution-free. Practitioners usually favor the less conservative BH, although the underlying statistical assumption is disputable. For example, in the interesting case of microarray data analysis, it is reported that the amplification of genes during the process can be very unequal as genes "compete" for the amount of polymerase available. A few RNA strands can "take over" early in the RT-PCR process, and, due to the exponential reaction, can let other strands non-amplified because of a lack of polymerase later in the process. Such an effect creates strong statistical dependencies between individual gene amplifications, in particular *negative* dependencies in the oberved expression levels.

This dicussion aside, we think there are several interesting added benefits in retrieving the BY procedure via Occam's hammer. First, in our opinion Occam's hammer sheds a totally new light on this kind of multi-testing procedure as the proof method followed in [3] was different and very specific to the framework and properties of statistical testing. Secondly, Occam's hammer allows us to generalize straightforwardly this procedure to an entire family by playing with the prior Π and more importantly the size prior Γ. In particular, it is clear that if something is known *a priori* over the expected size of the output, then this should be taken into account in the size prior Γ, possibly leading to a

more powerful testing procedure. Further, there is a significant hope that we can improve the accuracy of the procedure by considering priors depending on unknown quantities, but which can be suitably approximated in view of the data, thereby following the general principle of "self-bounding" algorithms that has proved to be quite powerful ([8], see also [5, 1] where this idea is used as well under a different form, called "localization"). This is certainly an exciting direction for future developments.

4 Tightness of Occam's hammer bound

It is of interest to know whether Occams' hammer is accurate in the sense that equality in the bound can be achieved in some (worst case) situations. A simple argument is that Occam's hammer is a generalization of Occam's razor: and since the razor is sharp [7], so is the hammer... This is somewhat unsatisfying since this ignores the situation Occam's hammer was designed for. In this section, we address this point by imposing an (almost) arbitrary inverse density prior ν and exhibiting an example where the bound is tight. Furthermore, in order to represent a "realistic" situation, we want the "bad sets" $B(h, \alpha)$ to be of the form $\{X_h > t(h, \alpha)\}$ where X_h is a certain real random variable associated to h. This is consistent with situations of interest described above (confidence intervals and hypothesis testing). We have the following result:

Proposition 4.1. *Let $\mathcal{H} = [0, 1]$ with interval extremities identified (i.e. the unit circumference circle). Let ν be a probability distribution on $[0, 1]$, and $\alpha_0 \in [0, 1]$ be given. Put $\beta(x) = \int_0^x u d\nu(u)$. Assume that β is a continuous, increasing function. Then there exists a family of real random variables $(X_h)_{h\in\mathcal{H}}$, having identical marginal distributions P and a random subset $A \subset [0, 1]$ such that, if $t(\alpha)$ is the upper α-quantile of P (i.e., $P(X > t(\alpha)) = \alpha$), then*

$$\mathbb{E}_{(X_h)}\left[\frac{|\{h \in A \text{ and } X_h > t(\alpha_0\beta(|A|))\}|}{|A|}\right] = \alpha_0 .$$

Furthermore, P can be made equal to any arbitrary distribution without atoms.

Comments. In the proposed construction (see proof in the appendix), the FPR is a.s. equal to α_0, and the marginal distribution of $|A|$ is precisely ν. This example shows that Occam's hammer can be sharp for the type of situation it was crafted for (set output procedures), and it reinforces the interpretation of ν as a "prior", since the bound is sharp precisely when the output distribution corresponds to the chosen prior. However, this example is still not entirely satisfying because in the above construction, we are basically oberving a single sample of (X_h), while in most interesting applications we have statistics based on averages of i.i.d. samples. If we could construct an example in which (X_h) is a Gaussian process, it would be fine, since observing an i.i.d. sample and taking the average would amount to a variance rescaling of the original process. In the above, although we can choose each X_h to have a marginal Gaussian distribution, the whole family is unfortunately not jointly Gaussian (inspecting the

proof, it appears that for $h \neq h'$ there is a nonzero probability that $X_h = X_{h'}$, as well as $X_h \neq X_{h'}$, so that $(X_h, X_{h'})$ cannot be jointly Gaussian). Finding a good sharpness example using a Gaussian process (the most natural candidate would be a stationary process on the circle with some specific spectral structure) is an interesting open problem.

5 Conclusion

We hope to have shown convincingly that Occam's hammer is a powerful and versatile theoretical device. It allows an alternate, and perhaps unexpected, approach to PAC-Bayes type bounds, as well as to multiple testing procedures. For the application to PAC-Bayes type bounds, an interesting feature of Occam's hammer approach is to provide a bound that is valid for the particular classifier returned by the randomization procedure and not just on average performance over the random output, and the former property is stronger. Furthermore, the tightest bounds available for a single classifier (i.e. by binomial tail inversion) can be plugged in without further ado. For multiple testing, the fact that we retrieve exactly the BY distribution-free multitesting procedure and extend it to a whole family shows that Occam's hammer has a strong potential for producing *practically useful* bounds and procedures. In particular, a very interesting direction for future research is to include in the priors knowledge about the typical behavior of the output set size. At any rate, a significant feat of Occam's hammer is to provide a strong first bridging between the worlds of learning theory and multiple hypothesis testing.

Finally, we want to underline once again that, like Occam's razor, Occam's hammer is a *meta* device that can apply on top of other bounds. This feature is particularly nice and leads us to expect that this tool will prove to have meaningful uses for other applications.

6 Appendix – additional proofs

Proof of Theorem 2.4. The proof of Occam's hammer is in essence an integration by parts argument, where the "parts" are level sets over $\mathcal{X} \times \mathcal{H}$ of the output density $\theta_X(h)$. We prove a slightly more general result than announced: let us consider a level function of the form

$$\Delta(h, \theta) = \min(\delta G(h, \theta^{-1}), 1)\,,$$

where $G : \mathcal{H} \times \mathbb{R}_+ \to \mathbb{R}_+$ is a measurable function which is nondecreasing in its second parameter, and satisfying

$$\int_{h \in \mathcal{H}} \int_{t \geq 0} G(h, t) t^{-2} d\Lambda(h) dt \leq 1\,.$$

Then the announced conclusion holds for this level function. First, note that the function $(X, h) \mapsto \mathbf{1}\{X \in \mathcal{B}(h, \Delta(h, \theta_X(h)))\}$ is jointly measurable in its two

variables by the composition rule using the measurability assumption in **(A)**; on $\theta_X(h)$ in the statement of the theorem; and on G above. We then have

$$\begin{aligned}
&\mathbb{P}_{X\sim P, h\sim\Theta_X}\left[X \in \mathcal{B}(h, \Delta(h, \theta_X(h)))\right] \\
&= \int_{(X,h)} \mathbf{1}\{X \in \mathcal{B}(h, \Delta(h, \theta_X(h)))\}\theta_X(h) d\Lambda(h) dP(X) \\
&= \int_{(X,h)} \mathbf{1}\{X \in \mathcal{B}(h, \Delta(h, \theta_X(h)))\} \int_{y>0} y^{-2}\mathbf{1}\{y \geq \theta_X(h)^{-1}\} dy dP(X) d\Lambda(h) \\
&= \int_{y>0} y^{-2} \int_{(X,h)} \mathbf{1}\{X \in \mathcal{B}(h, \Delta(h, \theta_X(h)))\}\mathbf{1}\{\theta_X(h) \geq y^{-1}\} dP(X) d\Lambda(h) dy \\
&\leq \int_{y>0} y^{-2} \int_{(X,h)} \mathbf{1}\{X \in \mathcal{B}(h, \Delta(h, y^{-1}))\} dP(x) d\Lambda(h) dy \\
&= \int_{y>0} y^{-2} \int_h \mathbb{P}_{X\sim P}\left[X \in \mathcal{B}(h, \min(\delta G(h,y), 1))\right] d\Lambda(h) dy \\
&\leq \int_{y=0}^{\infty} \int_h y^{-2}\delta G(h,y) d\Lambda(h) dy \leq \delta\,.
\end{aligned}$$

For the first inequality, we have used assumption **(A')** that $B(h,\delta)$ is an increasing family and the fact $\Delta(h,\theta)$ is a nonincreasing function in θ (by assumption on G). In the second inequality we have used the assumption on the probability of bad events. The other equalities are obtained using Fubini's theorem.

Now, it is easy to check that $G(h,t) = \pi(h)\beta(t)$ satisfies the above requirements, since it is obviously measurable, β is a nondecreasing function, and

$$\begin{aligned}
\int_{h\in\mathcal{H}} \int_{t\geq 0} \pi(h)\beta(t) t^{-2} d\Lambda(h) dt &= \int_h \pi(h) d\Lambda(h) \int_{t\geq 0}\int_{u\geq 0} u t^{-2}\mathbf{1}\{u \leq t\} d\Gamma(u) dt \\
&= \int_{u\geq 0} d\Gamma(u) = 1\,.
\end{aligned}$$

Note that in a more general case, if we have a joint prior probability distribution Γ on the product space $\mathcal{H}\times\mathbb{R}_+$, and if $\mathcal{H}$ is a Polish space, then there exists a regular conditional probability distribution $\Gamma(t|h)$, and the function $G(h,t) = \int_{u=0}^{t} u d\Gamma(u|h)$ is measurable and has the required properties by an obvious extension of the above argument. We opted to state our main result only in the case of a product prior for the sake of simplicity, but this generalization might be relevant for future applications. □

Proof of Corollary 3.2. Let us denote by $A_\delta \subset \mathcal{H}\times\mathcal{S}$ (here $\mathcal{S}$ denotes the set of samples S) the event where inequality (4) is violated; Proposition 3.1 states that $\mathbb{E}_{S\sim P}\left[\mathbb{P}_{h\sim\Theta_S}\left[(h,S)\in A_\delta\right]\right] \leq \delta$, hence by Markov's inequality, for any $\gamma \in (0,1)$ it holds with probability $1-\delta$ over the drawing of $S\sim P$ that

$$\mathbb{P}_{h\sim\Theta_S}\left[(h,S)\in A_{\delta\gamma}\right] \leq \gamma\,.$$

Let us consider the above statement for $(\delta_i, \gamma_i) = (\delta 2^{-i}, 2^{-i})$, and perform the union bound over the δ_i for integers $i \geq 1$. Since $\sum_{i\geq 1}\delta_i = \delta$, we obtain that

with probability $1-\delta$ over the drawing of $S \sim P$, it holds that for all integers $i \geq 0$ (the case $i = 0$ is trivial):

$$\mathbb{P}_{h\sim\Theta_S}\left[(h,S) \in A_{\delta 2^{-2i}}\right] \leq 2^{-i}\,.$$

From now on, consider a fixed sample S such that the above is satisfied. Let us denote

$$F(h,S) = nD_+(\widehat{\mathcal{E}}(h,S)\|\mathcal{E}(h_S)) - \log\left((k+1)\delta^{-1}\right) - \left(1+\frac{1}{k}\right)\log_+\theta_S(h)\,.$$

By the assumption on S, for all integers $i \geq 0 : \mathbb{P}_{h\sim\Theta_S}\left[F(h,S) \geq 2i\log 2\right] \leq 2^{-i}$; so that

$$\begin{aligned}\mathbb{E}_{h\sim\Theta_S}\left[F(h,S)\right] &\leq \int_{t>0}\mathbb{P}_{h\sim\Theta_S}\left[F(h,S)\geq t\right]dt \\ &\leq 2\log 2\sum_{i\geq 0}\mathbb{P}_{h\sim\Theta_S}\left[F(h,S)\geq 2i\log 2\right] \leq 3\,.\end{aligned}$$

Now let us detail some specific terms entering in the expectation $\mathbb{E}_{h\sim\Theta_S}\left[F(h,S)\right]$: we have

$$\mathbb{E}_{h_S\sim\Theta_S}\left[D_+(\widehat{\mathcal{E}}(h,S)\|\mathcal{E}(h_S))\right] \geq D_+\left(\mathbb{E}_{h_S\sim\Theta_S}\left[\widehat{\mathcal{E}}(h_S,S)\right]\middle\|\,\mathbb{E}_{h_S\sim\Theta_S}\left[\mathcal{E}(h_S)\right]\right)\,,$$

because the function D_+ is convex in is two joint parameters. Finally,

$$\begin{aligned}\mathbb{E}_{h_S\sim\Theta_S}\left[\log_+\theta_S(h)\right] &= \mathbb{E}_{h_S\sim\Lambda}\left[\theta_S(h)\log_+\theta_S(h)\right] \\ &\leq \mathbb{E}_{h_S\sim\Lambda}\left[\theta_S(h)\log\theta_S(h)\right] - \min_{0\leq x<1} x\log x \\ &= KL(\Theta_S\|\Lambda) + e^{-1}\,.\end{aligned}$$

Bounding e^{-1} by $1/2$ and gathering the terms leads to the conclusion. □

Proof of Proposition 4.1. Let ν and α_0 be fixed. We will construct explicitly the family $(X_h)_{h\in\mathcal{H}}$. Now, let U be a random variable uniformly distributed in $[0,1]$ and V an independent variable with distribution ν. We now define the family (X_h) given (U,V) the following way:

$$X_h = \begin{cases} g(V) & \text{if } h \in [U, U+\alpha_0 V]\,, \\ Y & \text{otherwise,}\end{cases}$$

where $g(v)$ is a decreasing real function $[0,1] \to [t_0,+\infty)$, and Y is a random variable independent of (U,V), and with values in $(-\infty,t_0]$. We will show that it is possible to choose g, Y, t_0 to satisfy the claim of the proposition. In the above construction, remember that since we are working on the circle, the interval $[U, U+\alpha_0 V]$ should be "wrapped around" if $U+\alpha_0 > 1$.

First, let us compute explicitly the quantile $t(\alpha)$ of X_h for $\alpha \leq \alpha_0$. We have assumed that $Y < t_0$ a.s., so that for any $h \in \mathcal{H}$, $t \geq t_0$,

$$\begin{aligned}\mathbb{P}\left[X_h > t\right] &= \mathbb{E}_u\left[\mathbb{P}\left[X_h > t|V\right]\right] = \mathbb{E}_V\left[\mathbb{P}\left[g(V) > t\,; h\in[U,U+\alpha_0 V]|V\right]\right] \\ &= \int_0^{g^{-1}(t)} \alpha_0 v d\nu(v) = \alpha_0\beta(g^{-1}(t))\,.\end{aligned}$$

Setting the above quantity equal to α, entails that $t(\alpha) = g(\beta^{-1}(\alpha_0^{-1}\alpha))$. Now, let us choose $A = [U, U+V]$ (note that due to the simplified structure of this example, the values of U and V can be inferred by looking at the family (X_h) alone since $[U, U+\alpha_0 V] = \{h : X_h \geq t_0\}$, hence A can really be seen as a function of the observed data alone). Then $|A| = V$, hence

$$t(\alpha_0\beta(|A|)) = g(\beta^{-1}(\alpha_0^{-1}\alpha_0\beta(V))) = g(V)\,.$$

This entails that we have precisely $A \cap \{h : X_h \geq t(\alpha_0(\beta(|A|)))\} = [U, U+\alpha_0 V]$, so that $|\{h \in A \text{ and } X_h \geq t(\alpha_0\beta(|A|)\}| \, |A|^{-1} = \alpha_0$ a.s. Finally, if we want a prescribed marginal distribution P for X_h, we can take t_0 as the upper α_0-quantile of P, Y a variable with distribution the conditional of $P(x)$ given $x < t_0$, and, since β is continuous increasing, we can choose g so that $t(\alpha)$ matches the upper quantiles of P for $\alpha \leq \alpha_0$. □

References

1. J.-Y. Audibert. Data-dependent generalization error bounds for (noisy) classification : a PAC-Bayesian approach. Technical Report PMA-905, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7, 2004.
2. Y. Benjamini and Y. Hochberg. Controlling the false discovery rate – a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B*, 57(1):289–300, 1995.
3. Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001.
4. A Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Occam's razor. *Information processing letters*, 24:377–380, 1987.
5. O. Catoni. A PAC-Bayesian approach to adaptive classification. Technical report, LPMA, Universtité Paris 6, 2004. (submitted to *Annals of Statistics*).
6. G. Gavin, S. Gelly, Y. Guermeur, S. Lallich, J. Mary, M. Sebag, and O. Teytaud. PASCAL theoretical challenge. Type I and type II errors for multiple simultaneous hypothesis testing. `http://www.lri.fr/~teytaud/risq`.
7. J. Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6:273–306, 2005.
8. J. Langford and A. Blum. Microchoice bounds and self bounding learning algorithms. *Machine Learning*, 51(2):165–179, 2003. (First communicated at COLT'99).
9. D. McAllester. Bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003. (First communicated at COLT'98 and '99).