Why Averaging Classifiers Can Protect Against Overfitting

Yoav Freund
AT&T Labs − Research
Shannon Laboratory
180 Park Avenue, Room A205
Florham Park, NJ 07932
yoav@research.att.com

Yishay Mansour*
Computer Science Dept.
Tel-Aviv University
ISRAEL
mansour@math.tau.ac.il

Robert E. Schapire
AT&T Labs − Research
Shannon Laboratory
180 Park Avenue, Room A203
Florham Park, NJ 07932
schapire@research.att.com

Abstract

We study a simple learning algorithm for binary classification. Instead of predicting with the best hypothesis in the hypothesis class, this algorithm predicts with a weighted average of all hypotheses, weighted exponentially with respect to their training error. We show that the prediction of this algorithm is much more stable than the prediction of an algorithm that predicts with the best hypothesis. By allowing the algorithm to abstain from predicting on some examples, we show that the predictions it makes when it does not abstain are very reliable. Finally, we show that the probability that the algorithm abstains is comparable to the generalization error of the best hypothesis in the class.

1 Introduction

Consider a binary classification learning problem. Suppose we use a hypothesis class $\mathcal{H}$ and are presented with a training set $(x_1, y_1), \ldots, (x_m, y_m)$ drawn independently from a distribution D over the example domain $X \times \{-1, +1\}$. Most learning algorithms for this problem that have been studied in computational learning theory are based on identifying the hypothesis $h \in \mathcal{H}$ that minimizes the training error. One of the main problems with this approach is the phenomenon called *overfitting*. Overfitting is encountered when the hypothesis class $\mathcal{H}$ is too "large," "complex" or "flexible" relative to the size of the training set. In this case, it is likely that the algorithm will find a hypothesis whose training error is very small but whose generalization error, or test error, is large. To overcome this problem, one usually uses either model-selection or regularization terms. Model selection methods try to identify the "right" complexity for $\mathcal{H}$. A regularization term is a measure of the complexity of the hypothesis h that is added to the training error to define a *cost* for each hypothesis. By minimizing this cost, the learning algorithm attempts to minimize both the training error and the amount of overfitting.

However, it is not clear that predicting with the hypothesis that minimizes the training error is indeed the only or the best thing to do. One popular alternative to predicting using the single best hypothesis is to *average* the prediction of those hypotheses whose performance on the training set is close to optimal. Two popular methods of this type are Bayesian averaging [9] and bagging [3, 4]. There is considerable experimental evidence that such averaging can significantly reduce the amount of overfitting suffered by the learning algorithm. However, there is, we believe, a lack of theory for explaining this reduction.

In the context of bagging, the common explanation is based on the argument that averaging reduces the variance of the classification rule. However, as argued elsewhere [6, 12], there is currently no adequate definition of variance for classification problems. In addition, this explanation fails to take into account the effect that the complexity of the model class has on overfitting.

In the Bayesian approach, the problem of overfitting is generally ignored. Instead, the basic argument is that the Bayesian method is always the best method, and therefore, the only important issues are how to choose a good prior distribution and how to efficiently calculate the posterior average. However, the optimality of the Bayesian method is based on the assumption that the data we observe are *generated* according to one of the distribution models *in the chosen class of models*. While this assumption is attractive for theory, it almost never holds in practice. In practice, one usually uses relatively simple models, either because there is not enough data to estimate the "true" model, because the computational complexity is prohibitive, or because our prior knowledge of the system is only partial. Even when very complex models are used, it is rarely the case that one can assume that the data are *generated* by a model in the class. As a result, Bayesian theory is inadequate for explaining why Bayesian prediction methods are better than predicting with the best model in the class.

In this paper, we propose a prediction method that is based on averaging among the empirically best classification rules. This method is similar to, but different from, the Bayesian method. The advantage of this method is that we can theoretically justify its usage without making the aforementioned Bayesian assumption that the data is generated by a distribution from a given class of distributions. Instead, we make the following weaker assumptions which are common in the context of empirical error minimization methods. First, we assume that the data is generated i.i.d. according to the distribution D defined above but make ab-

*Part of this work was done while visiting AT&T Labs.

solutely no assumption about D other than that it is a fixed distribution. Second, we choose a class of prediction rules (mappings from the input to the binary output) and assume that there are prediction rules in that class whose probability of error (with respect to the distribution D) is small, *but not necessarily equal to zero*.

We deviate from the analysis used for empirical error minimization methods in our definition of a classification *rule*. In the context of a binary prediction problem, we allow the classifier *three* possible outputs. Two of them, -1 and $+1$ are interpreted, as before, as predictions of the label. The third, denoted by 0, should be interpreted as "no prediction" or "insufficient data".

What is the benefit of allowing the predictor this new output? The advantage is that it allows the user of the classifier to identify those examples on which overfitting might occur. For example, suppose that the best hypothesis h in our hypothesis class $\mathcal{H}$ has an expected error of 1%. Suppose further that the size of the training set and the complexity of $\mathcal{H}$ are such that the hypothesis that minimizes the empirical error h^* is likely to have a generalization error of 5%. If we use h^* to make our predictions, then the most we can hope to get from a uniform-convergence type analysis is an upper bound on the generalization error that is close to 5%; we have no way of identifying *where* these errors might occur. On the other hand, if we allow the algorithm to output a zero, we can hope that the algorithm will output zero on about 4% of the input, and will be incorrect on about 1% of the data. In such a case, we say that the classifier *identifies* the locations of potential overfitting and allows the user to choose a special course of action for this case (such as referring the example back to a human to make the classification). In this case, we can justifiably say that the algorithm managed to avoid overfitting. It is not misleading us into thinking that we have a classifier that is very accurate just because its error on the training set is small.

As a toy example, Figure 1 shows a tiny learning problem in which positive and negative training examples are indicated by pluses and minuses. In this example, hypotheses are represented by rectangles, and we suppose that there is a large space of rectangular hypotheses, the best three of which are shown in the figure. Each of these makes two mistakes on this data set. However, if we take an average of hypotheses, one can imagine that it would be possible to obtain a combined classifier that abstains on all points in the shaded region where there is likely to be disagreement among the hypotheses, and predicts according to the weighted majority elsewhere. Such a combined classifier, when it does not abstain, would give nearly perfect predictions having successfully identified the regions where errors are most likely to occur.

Of course, if the generated classifier outputs zero most of the time, then there is no benefit from having it. We need to show two things to be convinced that the addition of the new output is useful. First, we need to show that the probability of outputting a zero is of the same order as the bounds on overfitting that we would get from an analysis based on uniform convergence. Second, we need to show that when the output is $+1$ or -1, the probability of making

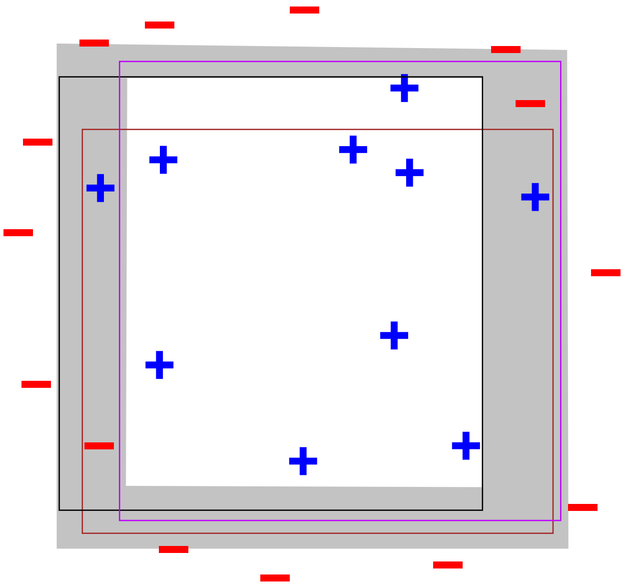

Figure 1: A toy example.

a mistake is similar to the generalization error of the best hypothesis in the class. In this paper, we prove that our algorithm has both these properties in the case that $\mathcal{H}$ is a finite class of models. In future work, we hope to show how this work can be extended to infinite model classes.

If $\mathcal{H}$ is finite, the uniform convergence bound is the well-known Occam's razor bound [2]. If $\mathcal{H}$ is infinite, we have to resort to bounds based on VC-dimension [14]. Unfortunately, these bounds are usually very loose and provide very poor estimates for the generalization error of learning algorithms in real-world applications.

In recent years, researchers in computational learning theory have started to consider algorithms that search for a good classification rule by optimizing quantities other than the training error. Algorithms of this type include support-vector machines [14] and boosting [12] which maximize the "margin" of a linear classifier. Other work by Shawe-Taylor and Williamson [13] and McAllester [10] provide PAC-style analyses of Bayesian algorithms. Bayesian algorithms compute the posterior distribution over the space of hypotheses and predict by averaging the predictions of all hypotheses whose training error is close to the minimum.

In this paper, we study a learning algorithm that is very similar to the algorithm that would be suggested by Bayesian analysis but uses a slightly different formula for computing the posterior distribution. This formula is the "exponential weights" formula introduced by Littlestone and Warmuth in the context of the weighted-majority algorithm [8] and further analyzed by Cesa-Bianchi et al. [5]. Note however that we are generating a fixed classification rule and are therefore working in the standard batch learning model and not in the online learning model.

The analysis of the algorithm consists of two parts. First, we consider, for each instance x, the log of the ratio of the total weight between those hypotheses that predict

$+1$ on x and those hypotheses that predict -1. We denote this ratio by $\hat{\ell}(x)$. We prove that $\hat{\ell}(x)$ is rather insensitive to the random choice of the training set. In particular, we prove that the variation in $\hat{\ell}(x)$ is *independent of the concept class* $\mathcal{H}$! This proof is interesting because it avoids using the standard "union bound"; in fact, it altogether avoids making any uniform claim on all of the hypotheses in $\mathcal{H}$.

Using this central theorem, we can show that if $\hat{\ell}(x)$ is far from zero, then predicting with $\text{sign}(\hat{\ell}(x))$ is very stable, i.e., is unlikely to change from training set to training set. More precisely, we introduce a non-stochastic quantity $\ell(x)$ and show that $\hat{\ell}(x)$ is, with high probability, very close to $\ell(x)$. Our algorithm predicts with $\text{sign}(\hat{\ell}(x))$ when $\hat{\ell}(x)$ is far from zero and abstains from prediction when $\hat{\ell}(x)$ is close to zero. We prove that the probability that this algorithm makes a prediction different from $\text{sign}(\ell(x))$ when it does not abstain is very small. On the other hand, we show that if $\mathcal{H}$ is finite and there is a hypothesis $\mathcal{H}$ whose error is ϵ then the error of $\text{sign}(\ell(x))$ is at most about 2ϵ.

The relation between our algorithm and algorithms that predict with the best hypothesis on the training set has a close correspondence to the relation between Bayesian prediction algorithms and MAP (maximum a-posteriori) algorithms. However, the analysis is carried out without making a Bayesian assumption, that is, we do not assume that the training data are generated by a model in a pre-specified class chosen by a pre-specified prior distribution. The prior and posterior distributions are internal to the algorithm and are not part of the world around it.

We hope that this paper will shed some new light on the use of algorithms that average many hypotheses such as Bayesian algorithms and averaging methods such as bagging [3, 4].

The paper is organized as follows. We start in Section 2 by describing the prediction algorithm. We give the basic analysis of the algorithm in Section 3. In Section 4, we bound the performance of $\ell(x)$ in terms of the error of the best hypothesis in the class. We conclude in Section 5 by giving a bound that is uniform with respect to the learning rate parameter η which makes it possible to choose this parameter after observing the training set.

2 The algorithm

Let D be a fixed but unknown distribution over (x, y) pairs, where $x \in X$ and $y \in \{-1, +1\}$. Let $\mathcal{H}$ be a fixed class of hypotheses, i.e., mappings from X to $\{-1, +1\}$. Let S denote a sample of m training examples, each drawn independently at random according to D. We denote the *true* error of a hypothesis h by $\varepsilon(h) \doteq \Pr_{(x,y)\sim D}\left[h(x) \neq y\right]$ and the estimated error according to the sample S by $\hat{\varepsilon}(h) \doteq \frac{1}{m}\sum_{i=1}^{m}\left[h(x) \neq y\right]$.

The prediction algorithm that we study calculates for each hypothesis h a *weight* that is defined as $w(h) \doteq e^{-\eta\hat{\varepsilon}(h)}$ where $\eta > 0$ is a parameter of the algorithm. The prediction on a new instance x is defined as a function of the *empirical log ratio*:

$$\hat{\ell}_\eta(x) \doteq \frac{1}{\eta}\ln\left(\frac{\sum_{h,h(x)=+1} w(h)}{\sum_{h,h(x)=-1} w(h)}\right)$$

$$= \frac{1}{\eta}\ln\left(\frac{\sum_{h,h(x)=+1} e^{-\eta\hat{\varepsilon}(h)}}{\sum_{h,h(x)=-1} e^{-\eta\hat{\varepsilon}(h)}}\right).$$

The prediction is defined to be

$$\hat{p}_{\eta,\Delta}(x) = \begin{cases} \text{sign}(\hat{\ell}(x)) & \text{if } |\hat{\ell}(x)| > \Delta \\ 0 & \text{otherwise} \end{cases}$$

where $\Delta \geq 0$ is a second parameter of the algorithm. Intuitively, the parameter Δ characterizes the range of values of $\hat{\ell}_\eta(x)$ in which the training data is insufficient to make a good prediction and a better choice is to abstain. When clear from context, we generally drop the subscripts and write simply $\hat{\ell}(x)$ and $\hat{p}(x)$.

3 Analysis of the algorithm

For an instance x, we define the *true log ratio* to be

$$\ell_\eta(x) \doteq \frac{1}{\eta}\ln\frac{\sum_{h,h(x)=+1} e^{-\eta\varepsilon(h)}}{\sum_{h,h(x)=-1} e^{-\eta\varepsilon(h)}}$$

which we often write as $\ell(x)$ when η is clear from context. The basic idea of our analysis is to show that $\hat{\ell}(x)$ must usually be close to $\ell(x)$ with high probability. In particular, we will prove the following two theorems. First, we will prove that for any fixed x the difference between the empirical log ratio and the true log ratio is small:

Theorem 1 *For any distribution D, any instance x, any $\lambda, \eta > 0$ and any $s \in \{-1, +1\}$:*

$$\Pr_{S\sim D^m}\left[s(\ell(x) - \hat{\ell}(x)) \geq 2\lambda + \frac{\eta}{8m}\right] \leq 2e^{-2\lambda^2 m}.$$

Then, in order to show that our algorithm has reasonable performance, we will transform Theorem 1 which gives a guarantee that holds with high probability for any *fixed* instance to a claim that holds with respect to a randomly chosen instance:

Theorem 2 *For any $\delta > 0$ and $\eta > 0$, if we set*

$$\Delta = 2\sqrt{\frac{\ln(\sqrt{2}/\delta)}{m}} + \frac{\eta}{8m}$$

then, with probability at least $1 - \delta$ over the random choice of the training set

$$\Pr_{(x,y)\sim D}\left[\hat{p}(x) \neq 0 \text{ and } \hat{p}(x) \neq \text{sign}(\ell(x))\right] \leq \delta.$$

This theorem guarantees that, when our algorithm predicts something different than 0 (which can be interpreted as "I don't know") it is very likely to be making the same prediction as $\ell(x)$. Note that the statements of Theorems 1 and 2 have no dependence on the hypothesis class $\mathcal{H}$. In fact, we believe the theorems and their proofs can be extended to infinite hypothesis classes, given the appropriate measurability assumptions.

We define some notation that will be used in the proofs. For $\mathcal{K} \subseteq \mathcal{H}$, let

$$R_\eta(\mathcal{K}) = \frac{1}{\eta}\ln\left(\sum_{h\in\mathcal{K}} e^{-\eta\varepsilon(h)}\right)$$

and let $\hat{R}_\eta(\mathcal{K})$ be the random variable

$$\hat{R}_\eta(\mathcal{K}) = \frac{1}{\eta} \ln \left(\sum_{h \in \mathcal{K}} e^{-\eta \hat{\varepsilon}(h)} \right).$$

We show that $\hat{R}_\eta(\mathcal{K})$ is close to $R_\eta(\mathcal{K})$ (with high probability) in two steps: First, we show that $\hat{R}_\eta(\mathcal{K})$ is close to its expectation $\mathbf{E}\big[\hat{R}_\eta(\mathcal{K})\big]$ with high probability. Then we show that $\mathbf{E}\big[\hat{R}_\eta(\mathcal{K})\big]$ is close to $R_\eta(\mathcal{K})$.

To prove the first result, we apply McDiarmid's theorem [11]:

Theorem 3 (McDiarmid) *Let $X_1, \ldots, X_m$ be independent random variables taking values in a set V. Let $f : V^m \to \mathbb{R}$ be such that, for $i = 1, \ldots, m$:*

$$|f(x_1, \ldots, x_m) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_m)| \leq c_i$$

for all $x_1, \ldots, x_m; x_i' \in V$. Then for $\epsilon > 0$, $s \in \{-1, +1\}$

$$\Pr\big[s\big(f(X_1, \ldots, X_m) - \mathbf{E}\big[f(X_1, \ldots, X_m)\big]\big) \geq \epsilon\big]$$

$$\leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^m c_i^2}\right).$$

Lemma 1 *Let $\mathcal{K}$ and $\hat{R}_\eta(\mathcal{K})$ be as above for a sample of size m. For $\eta > 0$, $\lambda > 0$ and $s \in \{-1, +1\}$*

$$\Pr\big[s\big(\hat{R}_\eta(\mathcal{K}) - \mathbf{E}\big[\hat{R}_\eta(\mathcal{K})\big]\big) \geq \lambda\big] \leq e^{-2\lambda^2 m}.$$

Proof: We apply McDiarmid's theorem with the X_i's set to the labeled examples of S, and the function f set equal to the random variable $\hat{R}_\eta(\mathcal{K})$. Let S' be the sample S in which one example (x_i, y_i) is replaced by (x_i', y_i'). Let $\hat{\varepsilon}'(h)$ be the empirical error of h on S', and let

$$\hat{R}'_\eta(\mathcal{K}) = \frac{1}{\eta} \ln \left(\sum_{h \in \mathcal{K}} e^{-\eta \hat{\varepsilon}'(h)} \right).$$

Then

$$\begin{aligned} \hat{R}'_\eta(\mathcal{K}) - \hat{R}_\eta(\mathcal{K}) &= \frac{1}{\eta} \ln \left(\frac{\sum_{h \in \mathcal{K}} e^{-\eta \hat{\varepsilon}'(h)}}{\sum_{h \in \mathcal{K}} e^{-\eta \hat{\varepsilon}(h)}} \right) \\ &\leq \frac{1}{\eta} \ln \left(\max_{h \in \mathcal{K}} e^{-\eta(\hat{\varepsilon}'(h) - \hat{\varepsilon}(h))} \right) \\ &= \max_{h \in \mathcal{K}} (\hat{\varepsilon}'(h) - \hat{\varepsilon}(h)) \leq \frac{1}{m}. \end{aligned}$$

The first inequality uses the fact that $(\sum_i a_i)/(\sum_i b_i) \leq \max_i a_i/b_i$ for positive a_i's and b_i's. The second inequality uses the fact that changing one example can change the empirical error by at most $1/m$.

By the symmetry of this argument, $|\hat{R}_\eta(\mathcal{K}) - \hat{R}'_\eta(\mathcal{K})| \leq 1/m$. Plugging in $c_i = 1/m$ in McDiarmid's theorem gives the result. ∎

Lemma 2 *Let $\mathcal{K}$, $R_\eta(\mathcal{K})$ and $\hat{R}_\eta(\mathcal{K})$ be as above for a sample of size m. Then for $\eta > 0$*

$$R_\eta(\mathcal{K}) \leq \mathbf{E}\big[\hat{R}_\eta(\mathcal{K})\big] \leq R_\eta(\mathcal{K}) + \frac{\eta}{8m}.$$

Proof: For the lower bound on $\mathbf{E}\big[\hat{R}_\eta(\mathcal{K})\big]$, let $\mathcal{K} = \{h_1, \ldots, h_N\}$, and let

$$g(x_1, \ldots, x_N) = \frac{1}{\eta} \ln \left(\sum_{i=1}^N e^{\eta x_i} \right).$$

Then g is convex. This can be seen by computing g's Hessian:

$$\frac{\partial^2 g}{\partial x_i^2} = C \left(e^{\eta x_i} \sum_{j=1}^N e^{\eta x_j} - e^{2\eta x_i} \right)$$

and

$$\frac{\partial^2 g}{\partial x_i \partial x_j} = -C e^{\eta(x_i + x_j)}$$

where

$$C = \frac{\eta}{\left(\sum_{i=1}^n e^{\eta x_i}\right)^2}.$$

Thus,

$$\frac{\partial^2 g}{\partial x_i^2} = \sum_{j \neq i} \left| \frac{\partial^2 g}{\partial x_i \partial x_j} \right|$$

which implies that g's Hessian matrix is diagonally dominant and therefore positive semidefinite. Hence, g is convex.

Therefore, by Jensen's inequality,

$$\begin{aligned} \mathbf{E}\big[\hat{R}_\eta(\mathcal{K})\big] &= \mathbf{E}\big[g(-\hat{\varepsilon}(h_1), \ldots, -\hat{\varepsilon}(h_N))\big] \\ &\geq g(-\mathbf{E}\big[\hat{\varepsilon}(h_1)\big], \ldots, -\mathbf{E}\big[\hat{\varepsilon}(h_N)\big]) \\ &= g(-\varepsilon(h_1), \ldots, -\varepsilon(h_N)) = R_\eta(\mathcal{K}). \end{aligned}$$

To prove the upper bound on $\mathbf{E}\big[\hat{R}_\eta(\mathcal{K})\big]$, we have by Jensen's inequality (applied to the concave log function),

$$\begin{aligned} \mathbf{E}\big[\hat{R}_\eta(\mathcal{K})\big] &= \frac{1}{\eta} \mathbf{E}\left[\ln \left(\sum_{h \in \mathcal{K}} e^{-\eta \hat{\varepsilon}(h)} \right) \right] \\ &\leq \frac{1}{\eta} \ln \left(\sum_{h \in \mathcal{K}} \mathbf{E}\left[e^{-\eta \hat{\varepsilon}(h)} \right] \right). \end{aligned} \tag{1}$$

Fix h and let $\varepsilon = \varepsilon(h)$ and $\hat{\varepsilon} = \hat{\varepsilon}(h)$. Let Z_i be a Bernoulli random variable that is 1 if $h(x_i) \neq y_i$ and 0 otherwise. Then we can write

$$\begin{aligned} \mathbf{E}\left[e^{\eta(\varepsilon - \hat{\varepsilon})} \right] &= \mathbf{E}\left[\exp\left(\frac{\eta}{m} \sum_{i=1}^m (\varepsilon - Z_i) \right) \right] \\ &= \prod_{i=1}^m \mathbf{E}\left[\exp\left(\frac{\eta}{m} (\varepsilon - Z_i) \right) \right] \\ &\leq \left(e^{\eta^2/8m^2} \right)^m = e^{\eta^2/8m}. \end{aligned}$$

The second equality uses independence of the Z_i's. The last step uses the fact, proved by Hoeffding [7], that for any random variable X with $\mathbf{E}[X] = 0$ and $a \leq X \leq b$, and for $s > 0$,

$$\mathbf{E}\big[e^{sX}\big] \leq e^{s^2(b-a)^2/8}.$$

Here we let $s = \eta/m$ and $X = \varepsilon - Z_i$.

Thus, $\mathbf{E}\left[e^{-\eta\hat{\varepsilon}(h)}\right] \leq e^{\eta^2/8m}e^{-\eta\varepsilon(h)}$. Combined with Eq. (1), this gives that

$$\mathbf{E}\left[\hat{R}_\eta(\mathcal{K})\right] \leq \frac{1}{\eta}\ln\left(e^{\eta^2/8m}\sum_{h\in\mathcal{K}} e^{-\eta\varepsilon(h)}\right) = R_\eta(\mathcal{K}) + \frac{\eta}{8m}$$

as claimed. ∎

Proof of Theorem 1: Given x, we partition the hypothesis set $\mathcal{H}$ into two. The subset $\mathcal{K}$ includes the hypotheses h such that $h(x) = +1$ and its complement $\mathcal{K}^c$ includes all h for which $h(x) = -1$. We can now write

$$\begin{aligned}\ell(x) - \hat{\ell}(x) &= \frac{1}{\eta}\ln\left(\frac{\sum_{h\in\mathcal{K}} e^{-\eta\varepsilon(h)}}{\sum_{h\in\mathcal{K}} e^{-\eta\hat{\varepsilon}(h)}}\right) \\ &\quad + \frac{1}{\eta}\ln\left(\frac{\sum_{h\in\mathcal{K}^c} e^{-\eta\hat{\varepsilon}(h)}}{\sum_{h\in\mathcal{K}^c} e^{-\eta\varepsilon(h)}}\right) \\ &= R_\eta(\mathcal{K}) - R_\eta(\mathcal{K}^c) - \hat{R}_\eta(\mathcal{K}) + \hat{R}_\eta(\mathcal{K}^c)\end{aligned} \tag{2}$$

Combining Lemma 1 and Lemma 2 we find that

$$\Pr\left[R_\eta(\mathcal{K}) - \hat{R}_\eta(\mathcal{K}) > \lambda\right] \leq e^{-2\lambda^2 m}. \tag{3}$$

and

$$\Pr\left[\hat{R}_\eta(\mathcal{K}^c) - R_\eta(\mathcal{K}^c) > \lambda + \frac{\eta}{8m}\right] \leq e^{-2\lambda^2 m}. \tag{4}$$

Combining Eqs. (2), (3) and (4) we prove the claim for $s = +1$. The proof for $s = -1$ is almost identical. ∎

Lemma 3 *For any distribution D, any $\lambda, \eta > 0$ and any $s \in \{-1, +1\}$, the probability over samples $S \sim D^m$ that*

$$\Pr_{(x,y)\sim D}\left[s(\ell(x) - \hat{\ell}(x)) \geq 2\lambda + \frac{\eta}{8m}\right] \geq \sqrt{2}e^{-\lambda^2 m}$$

is at most $\sqrt{2}e^{-\lambda^2 m}$.

Proof: Since Theorem 1 holds for all x, it also holds for a random x. Thus,

$$\begin{aligned}&\mathbf{E}_{S\sim D^m}\left[\Pr_{(x,y)\sim D}\left[s(\ell(x) - \hat{\ell}(x)) \geq 2\lambda + \frac{\eta}{8m}\right]\right] \\ &= \mathbf{E}_{(x,y)\sim D}\left[\Pr_{S\sim D^m}\left[s(\ell(x) - \hat{\ell}(x)) \geq 2\lambda + \frac{\eta}{8m}\right]\right] \\ &\leq 2e^{-2\lambda^2 m}.\end{aligned}$$

The lemma now follows using Markov's inequality. ∎

Theorem 2 follows immediately from this lemma.

4 Performance relative to the best hypothesis

We now show that there exists a setting of η and Δ that yields performance guarantees relative to the best hypothesis in the class. We compare these guarantees to those given by the Occam argument [2] for the algorithm that uses a hypothesis that minimizes the empirical error rate.

In Lemma 3, we showed that the value of $\hat{\ell}(x)$ is, with high probability, close to $\ell(x)$. We now show that, with respect to the *actual* distribution D, the sign of $\ell(x)$ is closely related to that of the best hypothesis in $\mathcal{H}$. By combining these theorems, we show that the generalization error of our algorithm is close to that of the best hypothesis in $\mathcal{H}$.

Note that the following theorem does not involve the training set in any way; it is a claim about $y\ell(x)$ which is a deterministic function of (x, y).

Theorem 4 *Let $\mathcal{H}$ be a finite hypothesis class and let ϵ be the error of the best hypothesis in $\mathcal{H}$ with respect to the distribution D over the examples. Let $\eta > 0$ and $\Delta \geq 0$ be such that $\Delta\eta \leq 1/2$. Then for any $\gamma \geq \ln(8|\mathcal{H}|)/\eta$,*

$$\Pr_{(x,y)\sim D}\left[y\ell(x) \leq 0\right] \leq 2\left(1 + 2|\mathcal{H}|e^{-\eta\gamma}\right)(\epsilon + \gamma),$$

and

$$\begin{aligned}&\Pr_{(x,y)\sim D}\left[y\ell(x) \leq 2\Delta\right] \\ &\leq \left(1 + e^{2\Delta\eta}\right)\left(1 + 2|\mathcal{H}|e^{\eta(2\Delta-\gamma)}\right)(\epsilon + \gamma) \\ &\leq 4\left(1 + 2|\mathcal{H}|e^{\eta(2\Delta-\gamma)}\right)(\epsilon + \gamma).\end{aligned}$$

Before proving the theorem, we give a corollary for a specific setting of the parameters η and Δ as a function of the sample size m, the size of the hypothesis class $\mathcal{H}$ and the reliability parameter δ.

Corollary 1 *Let $1/2 > \theta > 0$, $\delta > 0$ and*

$$\eta = \ln\left(8|\mathcal{H}|\right)m^{1/2-\theta};\ \Delta = 2\sqrt{\frac{\ln\left(\sqrt{2}/\delta\right)}{m}} + \frac{\ln\left(8|\mathcal{H}|\right)}{8m^{1/2+\theta}}.$$

For $m \geq 8$,

$$\Pr_{(x,y)\sim D}\left[y\ell(x) \leq 0\right] \leq \left(2 + \frac{1}{4m}\right)\left(\epsilon + \frac{\ln m}{m^{1/2-\theta}}\right),$$

and for

$$m \geq \left[8\sqrt{\ln\left(\frac{\sqrt{2}}{\delta}\right)}\ln(8|\mathcal{H}|)\right]^{1/\theta},$$

we have

$$\begin{aligned}&\Pr_{(x,y)\sim D}\left[y\ell(x) \leq 2\Delta\right] \\ &\leq 5\left(\epsilon + 4\sqrt{\frac{\ln\left(\sqrt{2}/\delta\right)}{m}} + \frac{\ln\left(8|\mathcal{H}|\right)}{4m^{1/2+\theta}} + \frac{1}{m^{1/2-\theta}}\right).\end{aligned}$$

Proof: To prove the corollary, we use Theorem 4 with two different settings of γ. The first bound is a result of choosing $\gamma = (\ln m)/m^{1/2-\theta}$, the second is a result of choosing $\gamma = 2\Delta + m^{\theta-1/2}$. ∎

We now discuss the significance of each statement in the corollary. Let us fix the reliability parameter δ.

The first statement of Corollary 1 shows that the sign of the true log ratio is a reasonably good proxy for the best hypothesis in the class. Specifically, the error of $\mathrm{sign}(\ell(x))$ is

$$2\varepsilon(h^*) + O\left(\frac{\ln(m)}{m^{1/2-\theta}}\right).$$

Combining this with the statement of Theorem 2, we find that the probability that our algorithm does not abstain and makes an incorrect prediction is upper bounded by

$$2\varepsilon(h^*) + O\left(\frac{\ln(m)}{m^{1/2-\theta}}\right) + \delta. \tag{5}$$

Note that this bound is *independent* of $|\mathcal{H}|$.

In comparison, the upper bound on the hypothesis that minimizes the empirical risk is

$$\varepsilon(h^*) + O\left(\sqrt{\frac{\ln(|\mathcal{H}|/\delta)}{m}}\right). \tag{6}$$

We see that the dependence on m here is slightly better, but the bound depends on the hypothesis class, which is what we expect from an algorithm that cannot abstain.

For our algorithm, the dependence on $|\mathcal{H}|$ instead appears in the bound on the probability of abstaining on a test example; this is given in the second statement of the corollary. Combining that statement with Lemma 3, we find that for

$$m = \Omega\left(\left(\sqrt{\ln(1/\delta)}\ln(|\mathcal{H}|)\right)^{1/\theta}\right),$$

our algorithm will predict zero with probability at most

$$5\varepsilon(h^*) + O\left(\frac{\sqrt{\ln(1/\delta)} + \ln(|\mathcal{H}|)}{m^{1/2-\theta}}\right).$$

This bound is similar to the Occam bound (Eq. 6), but the choice of θ makes an important difference in the dependence on m.

We now argue that the factor of two in front of the error of the best hypothesis in the class which appears in the first part of the corollary is necessary. Suppose that the example domain $X \times \{-1,+1\}$ is partitioned into two parts A_1 and A_2 such that $D(A_1) = 1 - 2\epsilon$ and $D(A_2) = 2\epsilon$. Suppose that all the hypotheses in $\mathcal{H}$ predict correctly on examples in A_1, while for each $(x,y) \in A_2$ each of the hypotheses predicts -1 or $+1$ with equal probability. In this case, each of the hypotheses in $\mathcal{H}$ has error about $1-\epsilon$; on the other hand, with high probability, our algorithm will predict 0 or "I don't know" on A_2 while it will always predict correctly on A_1.

In addition to showing that the factor of two has to be in the bound, this example points to the practical advantage of interpreting 0 as "I don't know" rather than as "$-1/+1$ with equal probability". The second interpretation would give us back a hypothesis that predicts correctly with probability $1-\epsilon$ but while masking the information that we have about the location of the prediction errors. On the other hand, the first interpretation retains this information so that we know when the predictions can be trusted and when they cannot.

At first, it may seem impossible that the bound in Eq. (5) is independent of the number of hypotheses. This would seem to suggest that overfitting can never happen, regardless of the complexity of the hypothesis space. In truth, if the hypothesis space is too complex, the algorithm will simply abstain more often. For example, suppose that the hypothesis space consists of *all* binary functions on a finite domain. For any set of training examples, there is a function that has zero training error (assuming no example appears twice with different labels). However, we expect any algorithm to be unable to predict the label of a new test example. Indeed, in this case, our algorithm will abstain on all unseen examples (since $\hat{\ell}(x)$ is exactly zero off the training set).

We now prove the theorem.

Proof of Theorem 4: We partition the hypotheses in $\mathcal{H}$ into two sets according to their true error. We call those hypotheses whose error is smaller than $\epsilon + \gamma$ *strong* and the other hypotheses *weak*.

We denote by W_w the total weight of the weak hypotheses:

$$W_w = \frac{1}{Z} \sum_{h \in \mathcal{H}:\ \varepsilon(h) \geq \epsilon+\gamma} e^{-\eta\varepsilon(h)}$$

where

$$Z = \sum_{h \in \mathcal{H}} e^{-\eta\varepsilon(h)}.$$

To upper bound W_w, note that we always have at least one strong hypothesis, namely, the one that achieves $\varepsilon(h) = \epsilon$. Thus,

$$W_w \leq \frac{|\mathcal{H}| e^{-\eta(\epsilon+\gamma)}}{e^{-\eta\epsilon}} = |\mathcal{H}| e^{-\eta\gamma}. \tag{7}$$

From the assumption that $\gamma \geq \ln(8|\mathcal{H}|)/\eta$, we get that $W_w \leq 1/8$.

For a given example (x,y), we partition the strong hypotheses into two subsets according to whether or not the hypothesis gives the correct prediction on (x,y). We denote the total weight of these subsets by

$$W_s^+(x,y) = \frac{1}{Z} \sum_{h \in \mathcal{H}:\ \varepsilon(h) < \epsilon+\gamma,\ h(x)=y} e^{-\eta\varepsilon(h)}$$

$$W_s^-(x,y) = \frac{1}{Z} \sum_{h \in \mathcal{H}:\ \varepsilon(h) < \epsilon+\gamma,\ h(x)\neq y} e^{-\eta\varepsilon(h)}.$$

By the definition of Z, for any (x,y),

$$W_s^+(x,y) + W_s^-(x,y) + W_w = 1.$$

We now prove the second part of the theorem; the first part follows from the second part by setting $\Delta = 0$. We first bound $y\ell(x)$ using W_w, $W_s^+(x,y)$ and $W_s^-(x,y)$:

$$y\ell(x) \geq \frac{1}{\eta} \ln\left(\frac{W_s^+(x,y)}{W_s^-(x,y) + W_w}\right).$$

Thus, $y\ell(x) \leq 2\Delta$ implies

$$\frac{W_s^-(x,y) + W_w}{1 - \left(W_s^-(x,y) + W_w\right)} \geq e^{-2\Delta\eta},$$

or equivalently,

$$W_s^-(x,y) + W_w \geq \frac{1}{1 + e^{2\Delta\eta}} \doteq c.$$

We denote by $h \sim \mathcal{S}$ the random choice of a hypothesis from the strong set with probability $e^{-\eta\varepsilon(h)}/Z_s$ where Z_s

normalizes the weights *within* the strong set to sum to 1. We find that

$$\begin{aligned}
&\Pr_{(x,y)\sim D}\left[y\ell(x) \le 2\Delta\right] \\
&\le \Pr_{(x,y)\sim D}\left[\frac{W_s^-(x,y)}{W_s^-(x,y)+W_s^+(x,y)} \ge \frac{c-W_w}{1-W_w}\right] \\
&= \Pr_{(x,y)\sim D}\left[\Pr_{h\sim S}\left[h(x)\neq y\right] \ge \frac{c-W_w}{1-W_w}\right] \\
&\le \mathbf{E}_{(x,y)\sim D}\left[\Pr_{h\sim S}\left[h(x)\neq y\right]\right]\frac{1-W_w}{c-W_w} \qquad (8) \\
&= \mathbf{E}_{h\sim S}\left[\Pr_{(x,y)\sim D}\left[h(x)\neq y\right]\right]\frac{1-W_w}{c-W_w} \qquad (9) \\
&\le (\epsilon+\gamma)\frac{1-W_w}{c-W_w} \qquad (10) \\
&\le (\epsilon+\gamma)\left(1+e^{2\Delta\eta}\right)\left(1+2W_w e^{2\Delta\eta}\right). \qquad (11)
\end{aligned}$$

Eqs. (8) and (9) use Markov's inequality and Fubini's theorem. Eq. (10) follows from the fact that $\varepsilon(h) < \epsilon + \gamma$ for every strong hypothesis. Eq. (11) uses our assumptions that $\Delta\eta \le 1/2$ and $W_w \le 1/8$ together with the inequality $(1-x)/(1-x(1+r)) \le 1 + 2xr$ for $x > 0$, $r > 0$ and $x(1+r) \le 1/2$ (with $x = W_w$ and $r = e^{2\Delta\eta}$).

Combining this bound with Eq. (7) proves the second statement of the theorem. ∎

Theorem 4 shows that the error of our predictor cannot be much worse than twice the error of the best hypothesis. On the other hand, it is possible in some favorable situations for our predictor to significantly outperform the best hypothesis. For example, suppose that there is an $h^* \in \mathcal{H}$ such that $\varepsilon(h^*) = 1/8$, and that for each $h \in \mathcal{H}' = \mathcal{H} - \{h^*\}$, we have $\varepsilon(h) = 1/4$. Suppose further that for each x, the fraction of $h \in \mathcal{H}'$ with the right label is $3/4$. Choosing the hypothesis with lowest observed error would give, hopefully, the hypothesis h^* that has an error rate of $1/8$. In our setting, for a labeled example (x, y), if $h^*(x) = y$, then

$$\begin{aligned}
y\ell(x) &= \frac{1}{\eta}\ln\left(\frac{e^{-\eta/8} + (3/4)|\mathcal{H}'|e^{-\eta/4}}{(1/4)|\mathcal{H}'|e^{-\eta/4}}\right) \\
&= \frac{1}{\eta}\ln\left(3 + \frac{4e^{\eta/8}}{|\mathcal{H}'|}\right).
\end{aligned}$$

Thus, for $\eta = 1$, we have $y\ell(x) = \ln(3+4e^{1/8}/|\mathcal{H}'|)$. Similarly, if $h^*(x) \neq y$ we have $y\ell(x) \ge \ln(3 - 12e^{1/8}/|\mathcal{H}'|)$. Note that this implies that $p_{1,0}(x)$ correctly classifies all the examples (for $|\mathcal{H}|$ large). Theorem 1, with λ set to a constant, then guarantees for $m = O(\lg 1/\delta)$ that $\hat{p}_{1,0}(x)$ has an error rate of at most δ. Note that in this example we choose to average (almost) uniformly the hypotheses although one hypothesis is clearly superior. In case there are more hypotheses with low error, the balance between the two sets becomes more delicate, and this is what our predictor performs.

5 Uniform bounds

The bound given in Lemma 1 applies to the case in which the parameter η is fixed ahead of time so that $\hat{R}_\eta(\mathcal{K})$ converges to $\mathbf{E}\left[\hat{R}_\eta(\mathcal{K})\right]$ for only a single value of η. In the next lemma, we show that on a single sample, this convergence is likely to take place for *all* values of $\eta \ge 1$ simultaneously.[1] The proof of this is primarily taken from Allwein, Schapire and Singer [1].

Lemma 4 *Let $\mathcal{K}$ and $\hat{R}_\eta(\mathcal{K})$ be as above for a sample of size m. For $\lambda > 0$,*

$$\begin{aligned}
&\Pr\left[\exists\eta \ge 1 : \left|\hat{R}_\eta(\mathcal{K}) - \mathbf{E}\left[\hat{R}_\eta(\mathcal{K})\right]\right| \ge \lambda\right] \\
&\qquad \le \frac{8\ln|\mathcal{K}|}{\lambda}e^{-\lambda^2 m/2}.
\end{aligned}$$

The proof is given in Appendix A.

We can now state the following theorems similar to Theorems 1 and 2. These theorems show that it is possible to design an algorithm that chooses η *after* the sample has been chosen without paying a large penalty in accuracy.

Theorem 5 *Let $\mathcal{K}$ and $\hat{R}_\eta(\mathcal{K})$ be as above for a sample of size m. For any distribution D, any $\lambda > 0$ and any $s \in \{-1, +1\}$:*

$$\begin{aligned}
&\Pr_{S\sim D^m}\left[\exists\eta \ge 1 \;:\; s(\ell_\eta(x) - \hat{\ell}_\eta(x)) \ge 2\lambda + \frac{\eta}{8m}\right] \\
&\qquad \le \frac{8\ln|\mathcal{K}|}{\lambda}e^{-\lambda^2 m/2}.
\end{aligned}$$

Theorem 6 *For any $\delta > 0$, if we set*

$$\Delta_\eta = 2\sqrt{\frac{2}{m}\ln\left(\frac{16m\ln|\mathcal{H}|}{\delta^2}\right)} + \frac{\eta}{m}$$

then, with probability at least $1-\delta$ over the random choice of the training set, for all $\eta \ge 1$

$$\Pr_{(x,y)\sim D}\left[\hat{p}_{\eta,\Delta_\eta}(x) \neq 0 \text{ and } \hat{p}_{\eta,\Delta_\eta}(x) \neq \mathrm{sign}(\ell_\eta(x))\right] \le \delta.$$

Acknowledgements

Thanks to Michael Cameron-Jones for helpful comments.

References

[1] Erin L. Allwein, Robert E. Schapire, and Yoram Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 9–16, 2000.

[2] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Occam's razor. *Information Processing Letters*, 24(6):377–380, April 1987.

[3] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[4] Leo Breiman. The heuristics of instability in model selection. *The Annals of Statistics*, 24:2350–2383, 1996.

[1] We can prove a similar result for $\eta > 0$ using a slightly more complicated proof. However, because η is typically large in this paper, we omit this proof.

[5] Nicolò Cesa-Bianchi, Yoav Freund, David Haussler, David P. Helmbold, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. *Journal of the Association for Computing Machinery*, 44(3):427–485, May 1997.

[6] Jerome H. Friedman. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1(1):55–77, 1997.

[7] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963.

[8] Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.

[9] David J. C. MacKay. *Bayesian Methods for Adaptive Models*. PhD thesis, California Institute of Technology, 1991.

[10] David A. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, 1998.

[11] Colin McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics 1989*, pages 148–188. Cambridge University Press, 1989.

[12] Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, October 1998.

[13] John Shawe-Taylor and Robert C. Williamson. A pac analysis of a bayesian estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, pages 2–9, 1997.

[14] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

A Proof of lemma 4

First, let $\mathcal{K} = \{h_1, \ldots, h_n\}$, and let

$$F(\eta, \mathbf{x}) = \frac{1}{\eta} \ln \left(\sum_{i=1}^{N} e^{-\eta x_i} \right).$$

For any $\mathbf{x}$, by checking derivatives, it can be verified that the function $\eta \mapsto F(\eta, \mathbf{x})$ is nonincreasing, while the function $\eta \mapsto F(\eta, \mathbf{x}) - (\ln N)/\eta$ is nondecreasing. Therefore, if $0 < \eta_1 \le \eta_2$ then for any $\mathbf{x} \in \mathbb{R}^N$,

$$0 \le F(\eta_1, \mathbf{x}) - F(\eta_2, \mathbf{x}) \le \left(\frac{1}{\eta_1} - \frac{1}{\eta_2} \right) \ln N. \qquad (12)$$

Now let

$$\mathcal{E} = \left\{ \frac{4 \ln N}{i\lambda} : i = 1, \ldots, \left\lfloor \frac{4 \ln N}{\lambda} \right\rfloor \right\}.$$

We show next that for any $\eta \ge 1$, there exists $\hat{\eta} \in \mathcal{E}$ such that

$$\left| \frac{1}{\eta} - \frac{1}{\hat{\eta}} \right| \ln N \le \frac{\lambda}{4}.$$

For if $\eta \ge 4(\ln N)/\lambda$ then let $\hat{\eta} = 4(\ln N)/\lambda$. Then

$$0 \le \left(\frac{1}{\hat{\eta}} - \frac{1}{\eta} \right) \ln N \le \frac{1}{\hat{\eta}} \ln N = \frac{\lambda}{4}.$$

Otherwise, if $1 \le \eta \le 4(\ln N)/\lambda$, then let $\hat{\eta} = 4(\ln N)/(i\lambda)$ be the smallest element of $\mathcal{E}$ that is no smaller than η. That is,

$$\frac{4 \ln N}{(i+1)\lambda} < \eta \le \frac{4 \ln N}{i\lambda}.$$

Then

$$\begin{aligned} 0 \le \left(\frac{1}{\eta} - \frac{1}{\hat{\eta}} \right) \ln N &= \left(\frac{1}{\eta} - \frac{i\lambda}{4 \ln N} \right) \ln N \\ &\le \left(\frac{(i+1)\lambda}{4 \ln N} - \frac{i\lambda}{4 \ln N} \right) \ln N \\ &= \frac{\lambda}{4}. \end{aligned}$$

Since $\hat{R}_\eta(\mathcal{K}) = F(\eta, \langle \hat{\varepsilon}(h_1), \ldots, \hat{\varepsilon}(h_N) \rangle)$, Eq. (12) and the argument above imply that for any $\eta \ge 1$, there exists $\hat{\eta} \in \mathcal{E}$ such that

$$\left| \hat{R}_\eta(\mathcal{K}) - \hat{R}_{\hat{\eta}}(\mathcal{K}) \right| \le \frac{\lambda}{4}$$

and so

$$\left| \left(\hat{R}_\eta(\mathcal{K}) - \mathbf{E}\left[\hat{R}_\eta(\mathcal{K}) \right] \right) - \left(\hat{R}_{\hat{\eta}}(\mathcal{K}) - \mathbf{E}\left[\hat{R}_{\hat{\eta}}(\mathcal{K}) \right] \right) \right| \le \frac{\lambda}{2}.$$

Thus,

$$\begin{aligned} &\Pr\left[\exists \eta \ge 1 : \left| \hat{R}_\eta(\mathcal{K}) - \mathbf{E}\left[\hat{R}_\eta(\mathcal{K}) \right] \right| \ge \lambda \right] \\ &\le \Pr\left[\exists \hat{\eta} \in \mathcal{E} : \left| \hat{R}_{\hat{\eta}}(\mathcal{K}) - \mathbf{E}\left[\hat{R}_{\hat{\eta}}(\mathcal{K}) \right] \right| \ge \frac{\lambda}{2} \right] \\ &\le 2|\mathcal{E}| e^{-\lambda^2 m/2} \end{aligned}$$

where the second inequality uses the union bound combined with Lemma 1. ∎