Predicting a binary sequence almost as well as the optimal biased coin

Yoav Freund

AT&T Laboratories
yoav@research.att.com
http://www.research.att.com/orgs/ssr/people/yoav/

April 3, 1996

Abstract

We apply the exponential weight algorithm, introduced and Littlestone and Warmuth [17] and by Vovk [24] to the problem of predicting a binary sequence almost as well as the best biased coin. We first show that for the case of the logarithmic loss, the derived algorithm is equivalent to the Bayes algorithm with Jeffrey's prior, that was studied by Xie and Barron under probabilistic assumptions [26]. We derive a uniform bound on the regret which holds for any sequence. We also show that if the empirical distribution of the sequence is bounded away from 0 and from 1, then, as the length of the sequence increases to infinity, the difference between this bound and a corresponding bound on the average case regret of the same algorithm (which is asymptotically optimal in that case) is only $1/2$. We show that this gap of $1/2$ is necessary by calculating the regret of the min-max optimal algorithm for this problem and showing that the asymptotic upper bound is tight. We also study the application of this algorithm to the square loss and show that the algorithm that is derived in this case is different from the Bayes algorithm and is better than it for prediction in the worst-case.

1 Introduction

In this paper we show how some methods developed in computational on-line learning theory can be applied to a very basic statistical inference problem and give what we think are surprisingly strong results. In order to present these results within context, we find it necessary to review some of the standard statistical methods used for this problem.

Consider the following very simple prediction problem. You observe a sequence of bits $x_1, x_2, \ldots$ one bit at a time. Before observing each bit you have to predict its value. The prediction of the tth bit x_t is given in terms of a number $p_t \in [0, 1]$. Outputting

p_t close to 1 corresponds to a confident prediction that $x_t = 1$ while p_t close to 0 corresponds to a confident prediction that $x_t = 0$. Outputting $p_t = 1/2$ corresponds to making a vacuous prediction.[1] Formally, we define a loss function $\ell(p, x)$ from $[0, 1] \times \{0, 1\}$ to the non-negative real numbers R^+. The value of $\ell(p_t, x_t)$ is the loss we associate with making the prediction p_t and then observing the bit x_t. We shall consider the following three loss functions: the *square loss* $\ell_2(p, x) = (x - p)^2$, the *log loss* $\ell_{\log}(p, x) = -x \log p - (1 - x) \log(1 - p)$ and the *absolute loss* $\ell_1(p, x) = |x - p|$.

The goal of the prediction algorithm is to make predictions that incur minimal loss. Of course, one has to make some assumption about the sequences in order to have any hope of making predictions that are better than predicting $1/2$ on all of the turns. Perhaps the most popular simple assumption is that the sequence is generated by independent random coin flips of a coin with some fixed bias p. The goal is to find an algorithm that minimizes the total loss incurred along the sequence. However, as the minimal achievable loss depends on p, it is more informative to consider the difference between the incurred loss and the minimal loss achievable by an "omniscient statistician" who knows the true value of p and chooses the optimal prediction $\tilde{p}$ for this distribution.

Let $\mathbf{x}^T = x_1, \ldots, x_T$ denote a binary sequence of length T, and let $\mathbf{x}^T \sim p^T$ denote the distribution of such sequences where each bit is chosen independently at random to be 1 with probability p. The *average regret* for the prediction algorithm A is the average difference between the total loss incurred by A and the total loss incurred by an omniscient statistician. In symbols, this is:[2]

$$R_{\text{av}}(A, T, p) \doteq E_{\mathbf{x}^T \sim p^T} \left(\sum_{t=1}^{T} \ell(p_t, x_t) - \sum_{t=1}^{T} \ell(\tilde{p}, x_t) \right) \quad (1)$$

We use the term *average regret* to differentiate it from the *worst-case regret* which we define below.

In general, the optimal algorithm for minimizing the average regret is the Bayes algorithm. There are two variants of the Bayesian methodology, we call these variants the *subjective Bayesianism* and the *logical Bayesianism*:

- **Subjective Bayesianism**
  In this view, the notion of a distribution over the set of models is defined, axiomatically, as a representation of the knowledge, or state of mind, of the statistician regarding the identity of the correct model in the model class. Choosing an appropriate prior distribution μ is, in this case, the act of compiling all such knowledge, that exists before seeing the data, into the form of a prior distribution. After μ has been chosen, the goal of a prediction algorithm is to minimize the expected average regret, where p is chosen at random according to the prior

[1] Strictly speaking, there exist non-symmetric loss function for which the vacuous prediction is not $1/2$. In this paper we concentrate on symmetric loss functions for which the vacuous prediction is always $1/2$.

[2] For the sake of simplicity, we restrict the discussion in the introduction to deterministic prediction algorithms. In this case the prediction is a function of the previously observed bits.

distribution and then $\mathbf{x}^T$ is generated according to p, i.e. to find A that minimizes $E_{p\sim\mu} R_{\text{av}}(A, T, p)$. It is easy to show that the Bayes algorithm with μ as a prior achieves this optimality criterion with respect to the log loss.

- **Logical Bayesianism**
  In this view, which is attributed to Wald, no assumption is made on how the model p is chosen, and the optimality criterion is that the average regret for the *worst-case choice* of p should be minimized. Interestingly, if the number of trials T is fixed ahead of time then the min-max optimal strategy for the adversary who chooses the value of p is to choose it according to some fixed distribution, $P_{\text{worst(T)}}$, over $[0, 1]$ (see e.g. Blackwell [3] Ferguson [11] and Haussler [13]). The min-max optimal prediction strategy for this case is the Bayes prediction algorithm with the prior set to $P_{\text{worst(T)}}$. However, these prior distributions are very peculiar (see e.g. [28]), calculating them is hard, and, most importantly, they are optimal only if T is known in advance. A more attractive option is to find an algorithm which does not need to know T in advance. Bernardo [2] suggested using the Bayes algorithm with Jeffrey's prior[3] (which we denote here by BJ), and Clarke and Barron [5] proved that this choice is asymptotically optimal for the models that are in the interior of the set. Finally, Xie and Barron [26] performed a detailed analysis of the BJ algorithm for the model class of biased coins and have shown that for any $\epsilon > 0$ and any $0 < \alpha < 1$:[4]

$$\lim_{T\to\infty} \left(\max_{\epsilon/T^\alpha \le p \le 1-\epsilon/T^\alpha} R_{\text{av}}(\text{BJ}, T, p) - \frac{1}{2} \ln T \right) = \frac{1}{2} \ln \frac{\pi}{2e} \tag{2}$$

These two methodologies, and most prediction methodologies in general, are based on the following statistical assumption. One assumes that the sequence to be predicted is *generated* by independent random draws from some unknown distribution that is selected, once and for all, from a known class of distributions. The difference between Bayesian and worst-case methodologies regards only the way in which the specific distribution is chosen from the class. However, the assumption that a particular source of sequences is really a so-called "random" source is problematic because in most real world cases it is almost impossible to verify that a particular data source is indeed a random source.[5] Moreover, because of the lack of data or computing power, one often wants to use a simple stochastic model even in situations where it is known that the sequence is generated by a much more complex random process or by a mechanism which is not random at all!

[3]Jeffrey's prior for this case is also known as the Krichevsky-Trofimov prior or the Dirichlet-$(1/2, 1/2)$ prior, see Equation (28) for the exact definition.

[4]Xie and Barron give their results in a slightly stronger form, our results can be presented in that form too.

[5]It seems that the most accepted approach to measuring the randomness of a particular sequence is to measure its Kolmogorov complexity (see e.g.[15]), in other words, to compare the length of the sequence with the length of the shortest program for generating it. Random sequences are those for which the program is not significantly shorter than that of the sequence. While this measure is mathematically very elegant, it is usually not a practical measure to compute.

The question is: what weaker formal assumption can be made that would still correspond to our intuition that a biased coin is a good approximate model for the data? The assumption we study in this paper is that *there exists a fixed model whose total loss on the sequence is non-trivial.* This direction builds upon the ideas of "prediction in the worst-case" [12], "on-line learning" [7, 16, 17, 24], "universal coding" [9, 22, 20], "universal portfolios" [8, 6] and "universal prediction" [10].

We define *trivial per-trial loss* as the maximal loss incurred by the best prediction. More formally, it is

$$\mathcal{L} \doteq \inf_{p\in[0,1]} \max_{x\in\{0,1\}} \ell(p,x) \; . \tag{3}$$

It is easy to verify that the prediction that minimizes the loss for the log loss, the square loss and the absolute loss is $1/2$ and that the corresponding values of the trivial loss are $\log 2$, $1/4$ and $1/2$ respectively. We say that a prediction algorithm makes non-trivial predictions on a particular sequence $x_1, \ldots, x_T$ if the total loss the algorithm incurs on the whole sequence is significantly smaller than $T\mathcal{L}$.

In the case of the biased coins, the assumption is that there exists some fixed value of $p \in [0,1]$ for which the total loss $\sum_{t=1}^{T} \ell(p, x_t)$ is significantly smaller than $T\mathcal{L}$. We do not define directly the concept of a "significant" difference. Instead, we define our requirements from the prediction algorithm relative to the total loss incurred by the optimal value of p. We can then say that ϵ is a significant difference in the total prediction loss if there exists a prediction algorithm whose total loss is guaranteed to be smaller than $T\mathcal{L}$ if $T\mathcal{L} - \sum_{t=1}^{T} \ell(p, x_t) > \epsilon$.

We measure the performance of a prediction algorithm on a specific sequence $x_1, \ldots, x_T$ using the difference

$$\sum_{t=1}^{T} \ell(p_t, x_t) - \min_{p\in[0,1]} \sum_{t=1}^{T} \ell(p, x_t) \; .$$

The *worst-case regret* of a prediction algorithm A is the maximum of this difference for sequences of length T is defined to be

$$R_{\mathrm{WC}}(A, T) \doteq \max_{\mathbf{x}^T} \left(\sum_{t=1}^{T} \ell(p_t, x_t) - \min_{p\in[0,1]} \sum_{t=1}^{T} \ell(p, x_t) \right) \tag{4}$$

We denote the argument that minimizes the second term by $\hat{p}$. Note that for the log loss and for the square loss $\hat{p}$ is equal to the fraction of 1's in $\mathbf{x}^T$. We refer to the fraction of 1's in $\mathbf{x}^T$ as the *empirical distribution* and denote it by $\hat{\theta}$. Clearly, $\hat{\theta}$ is always a rational number of the form i/T for some integer i in the range $0, \ldots, T$. For the absolute loss, $\hat{p}$ is 1 if $\hat{\theta} > 1/2$ and is 0 otherwise. As we shall see, we can prove slightly better bounds on the regret if we allow a dependence on $\hat{\theta}$, we therefor define the quantity

$$R_{\mathrm{WC}}(A, T, \hat{\theta}) \doteq \max_{\mathbf{x}^T;\, \hat{\theta}(\mathbf{x}^T)=\hat{\theta}} \left(\sum_{t=1}^{T} \ell(p_t, x_t) - \min_{p\in[0,1]} \sum_{t=1}^{T} \ell(p, x_t) \right) \; . \tag{5}$$

Consider the situation in which the sequence is generated by a random source and the analysis is done in terms of the worst-case regret. We have that

$$\begin{aligned}
& R_{\mathrm{WC}}(A,T) \\
= \; & \max_{\mathbf{x}^T} \max_{\hat{p}\in[0,1]} \sum_{t=1}^{T} \big(\ell(p_t,x_t) - \ell(\hat{p},x_t)\big) && (6)\\
\geq \; & \max_{p} E_{\mathbf{x}^T\sim p^T} \max_{\hat{p}\in[0,1]} \sum_{t=1}^{T} \big(\ell(p_t,x_t) - \ell(\hat{p},x_t)\big) && (7)\\
\geq \; & \max_{p} E_{\mathbf{x}^T\sim p^T} \sum_{t=1}^{T} \big(\ell(p_t,x_t) - \ell(p,x_t)\big) && (8)\\
= \; & \max_{p} R_{\mathrm{av}}(A,T,p) && (9)
\end{aligned}$$

The first inequality follows from replacing a maximum with an average and the second inequality follows from replacing the optimal per-sequence choice of $\hat{p}$ with a global choice of p. We thus see that the worst-case regret is an upper bound on the average-case regret. This relation is not very surprising. However, the surprising result, which we prove in this paper, is that for the class of biased coins and with respect to the log loss the worst case regret is only very slightly larger than the average case regret, as described in Equation (2). In this paper we prove the following bound on the worst case regret of the BJ algorithm for this problem. For any $\epsilon > 0$ and for any $0 < \alpha < 1$:

$$\lim_{T\to\infty} \left(\max_{\mathbf{x}^T;\, \hat{\theta}(\mathbf{x}^T)\in[\epsilon/T^\alpha,1-\epsilon/T^\alpha]} R_{\mathrm{WC}}(\mathrm{BJ},T) - \frac{1}{2}\ln T \right) \leq \frac{1}{2}\ln\frac{\pi}{2}\,. \qquad (10)$$

Observe that the difference between this bound and the bound given in Equation (2) is just $1/2$.

We also show that this tiny gap is necessary. We do this by calculating the min-max optimal algorithm for the worst-case regret and showing that its asymptotic difference from $(1/2)\ln T$ is also $(1/2)\ln(\pi/2)$. The only asymptotic advantage of the min/max algorithm is for sequences with empirical distribution very close to either 0 or 1, in which case the regret is larger by an additional $1/2$. Thus the algorithm suggested by the logical Bayesian analysis is also (almost) optimal with respect to the worst-case regret. This result complements the results of Xie and Barron [26].

This result also merges nicely with the method of stochastic complexity advocated by Rissanen [19]. That is because any prediction algorithm can be translated into a coding algorithm and vice versa (for example, using arithmetic coding [20].) The length of the code for a sequence $\mathbf{x}^T$ is equal (within one bit) to the cumulative log loss of the corresponding prediction algorithm on the same sequence. Thus our result means that the Bayes method using Jeffrey's prior is the optimal universal coding algorithm for the models class of biased coins, including the additive constant in the expression for the min-max redundancy, for all sequences but those whose empirical distribution is very extreme.

Thus, if our goal is to minimize the cumulative log loss or the total code length then stochastic complexity, logical Bayesianism and worst-case prediction all suggest using the same algorithm and all achieve essentially identical bounds. However, if we are interested in a loss function different than the log loss, then the worst-case methodology suggest an algorithm that differs significantly from the Bayes algorithm. Specifically, we show that the algorithm suggested for $\ell_2(p, x)$ is very different from the algorithm that is suggested by any Bayesian approach. Moreover, we give an example for which the worst-case regret of the Bayesian algorithm is significantly larger than that of the exponential weights algorithm.[6]

The prediction algorithms presented in this paper are very efficient. The prediction rule is a function only of t, the number of bits that have been observed, and n, the number of bits that were equal 1. The suggested prediction rule for the cumulative log loss is

$$p_t = \frac{t + 1/2}{n + 1} , \tag{11}$$

and a possible prediction rule for the square loss is

$$p_t = \sqrt{\frac{n^2}{t(t+1)} + \frac{1}{4} \ln \frac{t+1}{t} + \frac{1}{2} \ln \frac{\mathrm{erf}\left(\sqrt{\frac{2}{t}}n\right) + \mathrm{erf}\left(\sqrt{\frac{2}{t}}(t-n)\right)}{\mathrm{erf}\left(\sqrt{\frac{2}{t+1}}n\right) + \mathrm{erf}\left(\sqrt{\frac{2}{t+1}}(t+1-n)\right)}} \tag{12}$$

where

$$\mathrm{erf}\,(p) \doteq \frac{2}{\sqrt{\pi}} \int_0^p e^{-x^2} dx$$

is the cumulative distribution function of the normal distribution. Both prediction rules are close to $p_t = t/n$ for large n and t, but are slightly different for the initial part of the sequences.

The paper is organized as follows. In section 2 we review the exponential weights prediction algorithm and the well-known bound for the case where the number of models is finite. In Section 3 we show how the algorithm and its analysis can be extended to the case in which the class of models is uncountably infinite. In Section 4 we present our basic bound, that is based on the Laplace method of integration. In Section 5 we apply our method to the case of the log-loss and in Section 6 we compare our bound to other bounds regarding the cumulative log loss. In Section 7 we apply our method to the square loss and in Section 8 we briefly review what is known about the absolute loss. We conclude with some general comments and open problems in Section 9. Details of the proof are given in the appendix.

[6] It is a trivial observation that *any* prediction algorithm can be viewed as a Bayesian algorithm if the prior distribution is defined over the set of sequences, rather than over the set of models. However, this observation is of little value, because it does not suggest an interesting way for finding this distribution or for calculating the predictions that would be generated by using it.

2 The algorithm

The algorithm we study is a direct generalization of the "aggregating strategy" of Vovk [24, 23], and the "Weighted Majority" algorithm of Littlestone and Warmuth [17]. We refer to it as the "exponential weights" algorithm and denote it by EW. We denote a class of models by $\mathcal{P}$. In this section we assume that $\mathcal{P}$ is a finite set of values in $[0, 1]$ whose element are $p_1, \ldots, p_N$. We denote the cumulative loss of the model p_i at time t by $L(p_i, t) = \sum_{t'=1}^{t} \ell(p_i, x_{t'})$.

The algorithm is simple. It receives two positive real numbers η and c, as parameters. With each model in the class, at each time step t, it associates a *weight* $\omega_{t,i} = \exp\big(-\eta L(p_i, t)\big)$, The initial weights sum to 1, and, when the set of models is finite, the initial weights are usually set to $\omega_{1,i} = 1/N$ for all models $i = 1, \ldots, N$. The prediction of the algorithm at time t, which we denote by ϕ_t, is a function of the weights associated with the models at that time. Before describing how the predictions are chosen, we describe the bound on the total of the algorithm. This might seem backwards, but the choice of prediction is trivial once the bound is given. The bound on the total loss of the algorithm, for every time step t, is of the form

$$\sum_{t=1}^{T} \ell(\phi_t, x_t) \leq -c \ln \sum_{i=1}^{N} \omega_{T+1,i} \ . \tag{13}$$

If we want this bound to hold at all times for all sequences, it is clear how to make the predictions. The prediction should be chosen so that for both possible values of x_{t+1} the bound will hold at $t + 1$ given that it holds at t. Specifically, this means that the prediction ϕ_{t+1} is chosen so that

$$\begin{aligned} -c \ln \textstyle\sum_{i=1}^{N} \omega_{t,i} + \ell(\phi_t, 1) &\leq -c \ln \textstyle\sum_{i=1}^{N} \omega_{t+1,i} \\ \text{and } -c \ln \textstyle\sum_{i=1}^{N} \omega_{t,i} + \ell(\phi_t, 0) &\leq -c \ln \textstyle\sum_{i=1}^{N} \omega_{t+1,i} \ . \end{aligned} \tag{14}$$

The way this bound is usually used is to observe that if the total loss of the best model after observing all of $\mathbf{x}^T$ is $L_T^* \doteq \min_i L(p_i, T)$ then the weight associated with the best model, and thus also the total weight, are lower bounded by $\exp(-\eta L_T^*)$. Plugging this into Equation (13), we find, after some simple algebra, that

$$\sum_{t=1}^{T} \ell(\phi_t, x_t) \leq c \ln N + c\eta L_T^* \tag{15}$$

Thus if $c\eta = 1$ we immediately get a nice simple bound on the worst-case regret of the algorithm:

$$R_{\text{WC}}(E, t) \leq c \ln N \tag{16}$$

It is thus also clear that we would like $c = 1/\eta$ to be as small as possible.

Haussler, Kivinen and Warmuth [14] studied the problem of combining models for predicting a binary sequence in detail. They give a formula for calculating the minimal value of c for any loss function within a broad class such that the bound (13) holds for $\eta = 1/c$. In this paper we use their results for the log loss the square loss and the absolute loss.

3 Uncountably infinite sets of models

We now apply the exponential weights algorithm to the case where the set of models is the set of all biased coins. The natural extension of the notion of the weights that are associated with each model in a finite class is to assume that there is a *measure* defined over the set of models $\mathcal{P} = [0,1]$,[7] and as the initial weights sum to one, the initial measure, denoted by μ_1 is a probability measure. We shall sometimes refer to this initial probability measure as the *prior*. For $t > 1$ we define a measure $\mu_t(A)$ as follows:

$$\mu_{t+1}(A) \doteq \int_A e^{(1/c)l(x_t,p)} d\mu_t(p),$$

where $d\mu_t(p)$ denotes integration with respect to the measure μ_t over $p \in A \subseteq \mathcal{P}$. Similarly to the prediction rule given in Equation (14) the prediction at time t is any $\phi_t \in [0,1]$ which satisfies

$$\begin{aligned} l(0,\phi_t) &\leq -c \ln \left(\frac{\int_0^1 e^{-(1/c)l(0,p)} d\mu_t(p)}{\int_0^1 d\mu_t(p)} \right) \qquad (17) \\ \text{and } l(1,\phi_t) &\leq -c \ln \left(\frac{\int_0^1 e^{-(1/c)l(1,p)} d\mu_t(p)}{\int_0^1 d\mu_t(p)} \right) \end{aligned}$$

The bound that one is guaranteed in this case is

$$\sum_{t=1}^{T} l(x_t, \phi_t) \leq -c \ln \left(\int_0^1 d\mu_{T+1}(p) \right) \qquad (18)$$

Interestingly, the set of pairs (c, η) for which the bound of Equation (18) is guaranteed is identical to the set for which Equation (13) holds. The proofs are also identical, one has only to replace sums by integrals in the proofs given by Haussler et al. However, it is not immediately clear how to relate this bound on the total loss to the worst-case regret. As the number of models is infinite a bound of the form $c \log N$ is meaningless and we need a different bound. In the rest of the paper we develop a bound which is appropriate for the model class of the biased coins and is based on the method of Laplace integration.

[7]A measure over a space Ω is a function from the set of *measurable* sets (in our case, the Borel sets in $\Omega = [0,1]$) to the real numbers in the range $[0,1]$. A probability measure is a measure that assigns the value 1 to the set that consists of the whole domain.

From Equation (18) we get the following bound on the worst-case regret

$$R_{\mathrm{wc}}(\mathrm{EW}, T) \leq \max_{\hat{\theta}=i/T;\ i=0\ldots T} \left(-c\ln\left(\int_0^1 d\mu_{T+1}(p)\right) - L(\hat{\theta}, T)\right) \tag{19}$$

Notice now that in the case of the biased coin the cumulative loss of model p on the sequence $\mathbf{x}^T$ can be written in the form

$$L(p, T) = T\left[\hat{\theta}\ell(p, 1) + (1-\hat{\theta})\ell(p, 0)\right]$$

Using this expression for both $L(\hat{\theta}, T)$ and $\mu_{T+1}(p)$ and rewriting the second term (which is always positive) in an exponential form we get

$$R_{\mathrm{wc}}(\mathrm{EW}, T) \leq \max_{\hat{\theta}=i/T;\ i=0\ldots T} \left\{-c\ln\left(\int_0^1 \exp\left(-\frac{T}{c}\left[\hat{\theta}l(1, p) + (1-\hat{\theta})l(0, p)\right]\right) d\mu_1(p)\right)\right.$$
$$\left. -c\ln\exp\left(\frac{T}{c}\left[\hat{\theta}l(1, \hat{p}) + (1-\hat{\theta})l(0, \hat{p})\right]\right)\right\}$$

and we can combine the exponents of the two terms and get:

$$R_{\mathrm{wc}}(\mathrm{EW}, T) \leq \max_{\hat{\theta}=i/T;\ i=0\ldots T} -c\ln\left(\int_0^1 e^{-Tg(\hat{\theta}, p)} d\mu_1(p)\right) \tag{20}$$

where

$$g(\hat{\theta}, p) \doteq \frac{1}{c}\left[\hat{\theta}\left(\ell(p, 1) - \ell(\hat{p}, 1)\right) + (1-\hat{\theta})\left(\ell(p, 0) - \ell(\hat{p}, 0)\right)\right] \tag{21}$$

We refer to $g(\hat{\theta}, p)$ as the *gap function*. The gap function is proportional to the additional loss-per-trial that the model p suffers over and above the model $\hat{p}$ which is the optimal model for any sequence whose empirical distribution is $\hat{\theta}$. Thus the exponent of the integral in Equation 20 is zero when p is an optimal model and is negative elsewhere.

4 Laplace method of integration

In this section we describe a general method for calculating the integral in Equation (20). The derivation given in this section was done independently, for a much more general scenario, by Yamanishi in [27]. However, as we shall see, this method, by itself, is not sufficient to prove a bound on the worst-case regret. Later in this paper we describe the additional steps required to do that.

We require that the loss function $\ell(p, x)$ has the following three properties. It is easy to verify that the log loss and the square loss over the model class of biased coins have properties 2 and 3. The proof that property 1 holds for these loss functions can be found in Haussler et al [14].

1. $\ell(p, x)$ is $(c, 1/c)$ achievable for some $0 < x < \infty$. From here on we use the symbol c to denote the minimal value that satisfies this criterion.

2. For all values of x, $\ell(p, x)$ has a continuous second derivative as a function of p.

3. There exists a function $\hat{p} : [0, 1] \to [0, 1]$ such that the unique optimal model for any sequence $\mathbf{x}^T$ whose empirical distribution is $\hat{\theta}$ is $\hat{p}(\hat{\theta})$. We use $\hat{p}$ to denote $\hat{p}(\hat{\theta})$ when $\hat{\theta}$ is clear from the context.

The setting of the EW algorithm we suggest is to use the constants c and $\eta = 1/c$ defined in condition 1 and to use as the initial probability measure the following density measure:

$$\mu_1(A) = \int_A \omega(x) dx \tag{22}$$

$$\text{where} \quad \omega(x) = -\frac{1}{Z} \left[\frac{\partial^2}{\partial p^2} g(x, p) \right]_{p=\hat{p}(x)}$$

$$\text{and} \quad Z = \int_0^1 \left[\frac{\partial^2}{\partial p^2} g(x, p) \right]_{p=\hat{p}(x)} dx \tag{23}$$

The following theorem gives a bound on the performance of this algorithm:

Theorem 1 *For any fixed $0 < \hat{\theta} < 1$, the loss suffered by the exponential weights algorithm described above on any sequence $\mathbf{x}^T$ whose empirical distribution is $\hat{\theta}$ satisfies*

$$R_{WC}(A, T, \hat{\theta}) \leq \frac{c}{2} \ln \frac{T}{2\pi} - \frac{c}{2} \ln Z + O(1/T) , \tag{24}$$

where c and Z are as defined above.

This bound is *almost* a bound on the worst-case regret. However, it is an asymptotic result which applies only to sets of finite sequences in which all the sequences have the same empirical distribution, $\hat{\theta}$. Of course, any sequence has *some* empirical distribution, and so it belongs to some set of sequences for which the theorem holds. However, the term $O(1/T)$ might have a hidden dependence on $\hat{\theta}$.[8] What we need is a *uniform* bound, i.e. a bound that does not have *any* dependence on properties of the sequence. To get such a bound we need a more refined analysis which, at this point, we know how to do only for the special cases described in later sections. However, Theorem 1 is important because it bounds the regret for important sets of sequences, and because it suggests a choice for the initial probability measure.

The proof of Theorem 1 is based on the Laplace method of integration which is a method for approximating integrals of the form

$$\int_a^b f(t) e^{-Th(t)} dt , \tag{25}$$

[8] As we show later, such a dependence does indeed exist, but it vanishes as $T \to \infty$ if $p \in [\epsilon, 1 - \epsilon]$ for any fixed $\epsilon > 0$.

for large values of T, when f and h are sufficiently smooth functions from $[a, b]$ to the reals.[9] The intuition behind this method is that for large T the contribution of a small neighborhood of $t_{\min} \doteq \operatorname{argmin} g(t)$ dominates the integral. Thus, by using a Taylor expansion of $g(t)$ around $t = t_{\min}$ one can get a good estimate of the integral. The dependence of the contribution of the maximum on T depends on whether the maximum is also a point of derivative zero. This is always the case if $a < t_{\min} < b$ and might be the case if $t = a$ or $t = b$. We concentrate on the first case. Laplace method, or, more formally, Watson's Lemma [25], gives us the following asymptotic approximation for the integral in this case[10]

Theorem 2 (Watson) *Let f and h be functions from the segment $[a, b]$ to the reals. Assume that for all $t \in [a, b]$, $h(t) \geq 0$ and $\frac{d}{dt}h(t)$, $\frac{d^2}{dt^2}h(t)$ exist and are continuous. Assume also that there exists $a < t_{min} < b$ such that $h(t) = 0$ for $a \leq t \leq b$ if and only if $t = t_{min}$, and that $\left[\frac{d}{dt}h(t)\right]_{t=t_{min}} = 0$. Finally assume that $f(t)$ has a Taylor expansion in a neighborhood of t_{min}. Then*

$$\int_a^b f(t)e^{-Th(t)}dt = f(t_{min})\sqrt{\frac{-2\pi}{T\left[\frac{d^2}{dt^2}h(t)\right]_{t=t_{min}}}} + O(T^{-3/2}) \qquad (26)$$

We can now prove the theorem:

Proof of Theorem 1: We fix an empirical distribution $\hat{\theta}$ and let $\mathbf{x}^T$ be any sequence whose empirical distribution is $\hat{\theta}$. we use the fact the μ_1 is defined by the density function ω and rewrite the bound given in Equation (20), without taking the maximum over the sequence:

$$\sum_{t=1}^{T} \ell(p_t, x_t) - \sum_{t=1}^{T} \ell(p, x_t) \leq -c \ln\left(\int_0^1 \exp\left(-Tg(\hat{\theta}, p)\right)\omega(p)dp\right)$$

This integral is of the form defined in Equation (25), where $f(t) = \omega(t)/Z$, $h(t) = g(\hat{\theta}, t)$ and $t_{\min} = \hat{p}$. From Watson's Lemma we thus get that, for any fixed value of $\hat{\theta}$:

$$\int_0^1 e^{-Tg(\hat{\theta},p)}d\mu_1(p) = \omega(\hat{\theta})\sqrt{\frac{-2\pi c}{T\left[\frac{d^2}{dp^2}g(\hat{\theta}, p)\right]_{p=\hat{p}(\hat{\theta})}}} + O(T^{-3/2}) \,. \qquad (27)$$

In order to minimize the bound, we want the integral to be large. More precisely, we want to choose ω so as to maximize the minimal value achieved over all choices of $\hat{\theta} \in [0, 1]$.[11] As ω is a distribution, it is easy to see that the minimum of the first term

[9] For a good description of the Laplace method, see chapter 2 of Murray's book [18].

[10] See the derivation of Equation (2.33) in [18].

[11] Actually, if we fix T, we need to consider only values of $\hat{\theta}$ of the form i/T. Indeed, we can find slightly better choices of ω for fixed values of T. However, our goal is to choose a single distribution that will work well for all large T. We thus have to consider all rational $\hat{\theta}$, which, as the second derivative of h is continuous, is equivalent to considering all $\hat{\theta} \in [0, 1]$.

in the bound is maximized if the value of this term is equal for all values of $\hat{\theta}$. We thus arrive at the choice of ω given in Equation (22) which exactly cancels the dependence on $\hat{\theta}$ of the first term. We thus get

$$\max_{\hat{\theta}\in[0,1]} \omega(\hat{\theta}) \sqrt{\frac{-2\pi c}{T\left[\frac{d^2}{dp^2} g(\hat{\theta}, p)\right]_{p=\hat{p}(\hat{\theta})}}} + O(T^{-3/2}) \leq \sqrt{\frac{2\pi Z}{T}} + O(T^{-3/2}) = \sqrt{\frac{2\pi Z}{T}}\left(1 + O(1/T)\right)$$

and when we plug this estimate into the bound given in Equation 20 we get the statement of the theorem. ■

We now move on to show that the suggested exponential weights algorithm does indeed achieve a very strong bound on the worst-case regret for the log loss and for the square loss on the class of biased coins.

5 Log-loss

The loss function in this case is $\ell_{\log}((,x),p) = -\log(1-|x-p|)$, the optimal parameters are $c = 1/\eta = 1$, and the optimal value of p for a given sequence is $\hat{p} = \hat{\theta}$.

It is easy to check that in this case the prediction rule

$$\phi_t = \int_0^1 p \;\; \exp\left(-T\left(\hat{\theta}\ln\hat{\theta} + (1-\hat{\theta})\ln(1-\hat{\theta})\right)\right) d\mu_1(p),$$

satisfies Equation (14) for the log loss. Note also that this rule is equivalent to the Bayes optimal prediction rule using the prior distribution μ_1.

The gap function h in this case is (minus) the KL-divergence.

$$g(\hat{\theta}, p) = -\hat{\theta}\log\left(\frac{\hat{\theta}}{p}\right) - (1-\hat{\theta})\log\left(\frac{1-\hat{\theta}}{1-p}\right) = -D_{\text{KL}}\left(\hat{\theta}||p\right)$$

The second derivative of h is the Fisher information:

$$\left[\frac{\partial^2}{\partial p^2} g(x,p)\right]_{p=\hat{p}(x)} = p(1-p)$$

So the optimal prior is

$$\omega(p) = \frac{1}{Z\sqrt{\left[\frac{\partial^2}{\partial p^2} g(x,p)\right]_{p=\hat{p}(x)}}} = \frac{1}{\pi\sqrt{p(1-p)}}. \tag{28}$$

This prior is the Jeffrey's prior for this model class, thus the algorithm suggested in this case is the Bayes algorithm using Jeffrey's prior (BJ).

To bound the worst-case regret we calculate the integral of Equation (18) which, in this case, is equal to

$$\int_0^1 \omega(p) e^{-T D_{\text{KL}}(\hat{\theta}||p)} dp$$

Details of this calculation will be given in appendix A. Here we state the resulting bound:

Theorem 3 *The regret of the Exponential weights algorithm over the class of biased coins, which uses the prior distribution described in Equation (28) with respect to the log loss is bounded, for any* $T \geq 1$ *by*

$$R_{WC}(EW, T, \hat{\theta}) \leq \tag{29}$$
$$\frac{1}{2}\ln(T+1) + \frac{1}{2}\ln\frac{\pi}{2} + \frac{1}{2} - \frac{1}{2 + (T\min(\hat{\theta}, 1-\hat{\theta}))^{-1}} + \frac{1}{360(T+1)^3}$$

which implies that

$$R_{WC}(EW, T) \leq \frac{1}{2}\ln(T+1) + 1 \tag{30}$$

The last inequality shows that the regret of the exponential weights algorithm holds uniformly for all $\hat{\theta} \in [0,1]$, i.e., for all sequences. It is worthwhile to consider the more precise bound given in Inequality (29). If for some fixed $\epsilon > 0$ we have that $\hat{\theta} \in [\epsilon, 1-\epsilon]$ then, for $T \to \infty$ the last two terms converge to zero and the bound converges to $\frac{1}{2}\ln T + \frac{1}{2}\ln\frac{\pi}{2}$. This bound also holds if $\hat{\theta} \in [\frac{\epsilon}{T^\alpha}, 1 - \frac{\epsilon}{T^\alpha}]$ for any $\alpha < 1$. However, if $\hat{\theta} = 0$, we get a slightly larger bound of $\frac{1}{2}\ln T + \frac{1}{2}\ln\frac{\pi}{2} + \frac{1}{2}$.[12] Finally, if $\hat{\theta} = \Theta(1/T)$ then we get an intermediate bound.

6 Comparison to other results regarding log-loss

It is interesting to compare these bounds to the ones given by Xie and Barron [26]. They analyze the same algorithm on a very similar problem, but they consider the expected regret and not the worst-case regret. As was shown in the introduction, the worst-case regret upper bounds the average-case regret. However, our definition of regret is much stronger then theirs, because we make no probabilistic assumption about the mechanism that is generating the sequence. It is therefor surprising that the bounds that we get are so very close to their bounds.

From the arguments given in the previous section we get that, Theorem 3 implies the bound given in Equation (10). The difference between this bound and the bound derived by Xie and Barron [26] described in Equation (2) is $(1/2)\ln(\pi/2) - (1/2)\ln(\pi/2e) =$ 0.5 nits $= 0.721$ bits . In other words, knowing that the sequence is actually generated by independent random draws of a random coin is worth less than one bit!

As Xie and Barron show, the BJ algorithm is not an asymptotically optimal algorithm with respect to the average prior. That is because on the endpoints $p = 0$ and $p = 1$ the loss is larger than for the interior points, and this difference does not vanish as $T \to \infty$. In order to achieve the asymptotic min/max they suggest multiplying Jeffrey's prior by $1 - 2\eta$ where $\eta > O(T^{-\alpha})$ for some $\alpha > 1/2$ and putting a probability mass of

[12] The actual asymptotic value of the regret (both worst case and average case) of the BJ algorithm for $\hat{\theta} = 0$ or $\hat{\theta} = 1$ is slightly smaller: $\frac{1}{2}\ln T + \frac{1}{2}\ln \pi$.

η on each of the two points c/T and $1 - c/T$ for some constant c. This reduces the asymptotic average regret at the endpoints to the asymptotically optimal value without changing the asymptotic average regret in the interior points. Similar observations and the same fix hold for the worst case regret.

As we show at the end of the last section, the regret of algorithm BJ on the interior of $[0, 1]$ is $(1/2)\ln T + (1/2)\ln(\pi/2)$, larger by $1/2$ from the optimal performance with respect to the average regret. We now show that this small gap cannot be removed. We do this by calculating the regret of the min-max optimal algorithm for the worst-case regret with respect to the log loss.

We use a rather well known lemma, stated for instance in Starkov [21, 22] and in Cesa-Bianchi et al. [4]. The lemma states that the min/max optimal prediction algorithm defines the following distribution over the set of sequences of length T:

$$P(\mathbf{x}^T) = \frac{1}{Z}\max_p P_p(\mathbf{x}^T) \quad \text{where} \quad Z = \sum_{\mathbf{x}^T} \max_p P_p(\mathbf{x}^T)\ , \tag{31}$$

and that the regret suffered by this optimal algorithm on any sequence is equal to $\ln Z$. Using this we can explicitly calculate the worst-case regret of the min/max optimal algorithm (this result was previously shown by Starkov [22]).

Lemma 1 *The worst-case regret of the min/max optimal prediction algorithm for sequences of length T, which we denote by MM_T, with respect to the class of biased coins and the log loss, is*

$$\begin{aligned} R_{WC}(MM_T, T) &= \ln\left(\sum_{i=0}^{T}\binom{T}{i} e^{-T\ H(i/T)}\right) \\ &= \frac{1}{2}\ln(T+1) + \frac{1}{2}\ln(\pi/2) - O(1/\sqrt{T}) \end{aligned} \tag{32}$$

The proof is given in Appendix B.

7 Square loss

In this section we consider the loss function $l(x,p) = (x-p)^2$. As was shown by Vovk [24] and Haussler et al. the optimal parameters in this case are $c = 1/2, \eta = 2$ and the optimal model is $\hat{p} = \hat{\theta}$. The gap function in this case is

$$g(p) = \hat{\theta}\left((1-\hat{\theta})^2 - (1-p)^2\right) + (1-\hat{\theta})\left((0-\hat{\theta})^2 - (0-p)^2\right)$$

And its second derivative is a constant:

$$h''(p) = 4$$

So the optimal prior is the uniform distribution

$$\omega(p) = 1$$

To bound the worst-case regret we calculate the integral of Equation (18) which, in this case, is equal to

$$\int_0^1 \exp\left(2T\hat{\theta}\left((1-\hat{\theta})^2-(1-p)^2\right) + 2T(1-\hat{\theta})\left((0-\hat{\theta})^2-(0-p)^2\right)\right) dp$$

Details are given in appendix C. The resulting bound is:

Theorem 4 *The regret of the Exponential weights algorithm over the class of biased coins, which uses the uniform prior distribution with respect to the the squared loss, for any $T \geq 1$, is bounded by*

$$R_{WC}(EW, T, \hat{\theta}) \leq \frac{1}{4}\ln T + \frac{1}{2}\ln\frac{2}{erf\left(\hat{\theta}\sqrt{T}\right) + erf\left((1-\hat{\theta})\sqrt{T}\right)} - \frac{1}{4}\ln\frac{\pi}{2} \quad (33)$$

which implies

$$R_{WC}(EW, T) \leq \frac{1}{4}\ln T + \frac{1}{2}\ln\frac{2}{erf\left(\sqrt{2}\right)} - \frac{1}{4}\ln\frac{\pi}{2} \quad (34)$$

Similar to the detailed analysis of the log-loss case, Inequality (34) gives us a uniform upper bound that does not depend on the sequence, while if we assume that $\hat{\theta} \in [\epsilon, 1-\epsilon]$ for some constant $\epsilon > 0$ then Inequality (33) gives a slightly better asymptotic bound. In the second case each of the two $\mathrm{erf}(\cdot)$ functions converges to 1 and so the second term vanishes and we are left with the negative term $-(1/4)\ln(\pi/2)$. If $\hat{\theta} = 0$ or $\hat{\theta} = 1$ then only one of the two $\mathrm{erf}(\cdot)$ terms converges to 1 while the other remains 0, and we get an additional term of $\ln(2)/2$ in the regret. For the square loss the restriction on the distance between $\hat{\theta}$ and 0 (or 1) is a bit stronger then in the log-loss case. Here we have that if $\hat{\theta} \in [\epsilon/T^{\alpha}, 1-\epsilon/T^{\alpha}]$ for some $\alpha < 1/2$ then the better bound holds and if $\hat{\theta} = \Theta(1/\sqrt{T})$ then we get a bound that is between the interior bound and the bound for $\hat{\theta} = 0$.

Two comments are in order. First, although we have not concerned ourselves with the computational efficiency of our algorithms, both the log-loss version and the square-loss version require small constant time to calculate the predictions, whose formulas are given in Equations 11 and 12. Second, it is not hard to give examples of finite model classes in which using the EW algorithm is much better than using any Bayes algorithm when the data is generated by a model outside the class (See Appendix D). We conjecture that such an example exists also for the continuous model class $\mathcal{P} = [0, 1]$.

8 Absolute Loss

In this section we consider the absolute loss, which has very different properties than the log loss and the square loss. Haussler et al. [14] there is no finite value of c such that the bound (13) holds for $\eta = 1/c$. Moreover, as was shown by Cesa-Bianchi at al. [4], there is no prediction algorithm whose worst-case regret does not depend on the loss of the best model. On the other hand, there are choices of c and $\eta > 1/c$ for which Equation (13) holds, and Cesa-Bianchi et. al. have shown that, based on this fact, an exponential weights algorithm for finite model classes, can be devised. And the worst-case regret of this algorithm is bounded by $O(\sqrt{\ln N L_T^*})$.

As we cannot choose c finite and $\eta = 1/c$ for the multiplicative weights algorithm, we cannot use the technique of Theorem 1 for this case. However, we do not need to use an infinite set of models in our analysis for this case. This is because in this case the optimal model in the class $\mathcal{P} = [0, 1]$ is always either $p = 0$ or $p = 1$. Thus we can consider a EW algorithm that combines only these two models and get close to optimal bounds on the regret.

9 Conclusions and open problems

We have demonstrated, in a simple case, that the Bayes algorithm that has been shown by Xie and Barron to be optimal with respect to the average-case regret is also optimal with respect to the worst-case regret. Moreover, the bound on the worst-case regret is only very slightly worse than the average-case regret.

We have also shown that a very different algorithm results if one is interested in the square loss, rather than in the log loss.

These results give evidence that sometimes accurate statistical inference can be done without assuming that the world is random. We are currently working on extended this work to more general classes of models of sequences over larger alphabets.

Acknowledgments

Special thanks to Sebastian Seung for his help on understanding and solving the integral equations used in this paper. Thanks to Andrew Barron, Meir Feder, David Haussler, Rob Schapire, Volodya Vovk, Qun Xie and Kenji Yamanishi for helpful discussion and suggestions.

References

[1] Abramowitz and Stegun. *Handbook of Mathematical Functions*. National Bureau of Standards, 1970.

[2] J. M. Bernardo. Reference posterior distributions for bayesian inference. *J. Roy. Statistic. Soc. Ser. B.*, 41:113–147, 1979.

[3] David Blackwell and M.A. Girshick. *Theory of Games and Statistical Decisions*. Dover Publications, Inc, New York, 1954.

[4] Nicolò Cesa-Bianchi, Yoav Freund, David P. Helmbold, David Haussler, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. In *Proceedings of the Twenty-Fifth Annual ACM Symposium on the Theory of Computing*, pages 382–391, 1993. To appear, *Journal of the Association for Computing Machinery*.

[5] Bertrand S. Clarke and Andrew R. Barron. Jeffrey's prior is asymptotically least favorable under entropic risk. *J. Stat Planning and Inference*, 41:37–60, 1994.

[6] T. Cover and E. Ordentlich. Universal portfolios with side information. Unpublished Manuscript, 1995.

[7] Thomas M. Cover. Behavior of sequential predictors of binary sequences.

[8] Thomas M. Cover. Universal portfolios. *Mathematical Finance*, 1(1):1–29, January 1991.

[9] L. D. Davisson. Universal noiseless coding. *IEEE Trans. Inform. Theory*, 19:783–795, 1973.

[10] M. Feder, N. Merhav, and M. Gutman. Universal prediction of individual sequences. *IEEE Transactions on Information Theory*, 38:1258–1270, 1992.

[11] Thomas S. Ferguson. *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, 1967.

[12] Dean P. Foster. Prediction in the worst case. *The Annals of Statistics*, 19(2):1084–1090, 1991.

[13] David Haussler. A general minimax result for relative entropy. Unpublished manuscript.

[14] David Haussler, Jyrki Kivinen, and Manfred K. Warmuth. Tight worst-case loss bounds for predicting with expert advice. In *Computational Learning Theory: Second European Conference, EuroCOLT '95*, pages 69–83. Springer-Verlag, 1995.

[15] Ming Li and Paul Vitányi. *An introduction to Kolmogorov complexity and its applications*. Texts and Monogaraphs in Computer Science. Springer-Verlag, 1993.

[16] Nick Littlestone. Learning when irrelevant attributes abound. In *28th Annual Symposium on Foundations of Computer Science*, pages 68–77, October 1987.

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

[18] J.D. Murray. *Asymptotic Analysis*. Springer Verlag, 1973.

[19] Jorma Rissanen. *Stochastic Complexity in Statistical Inquiry*, volume 15 of *Series in Computer Science*. World Scientific, 1989.

[20] Jorma Rissanen and Glen G. Langdon, Jr. Universal modeling and coding. *IEEE Transactions on Information Theory*, IT-27(1):12–23, January 1981.

[21] Yu. M. Shtarkov. Coding of descrete sources with unknown statistics. In I. Csiszar and P. Elias, editors, *Topics in Information Theory*, pages 559–574. North Holland, Amsterdam, 1975.

[22] Yu. M. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23:175–186, July-September 1987.

[23] V. G. Vovk. A game of prediction with expert advice. In *Proceedings of the Eighth Annual Conference on Computational Learning Theory*, 1995.

[24] Volodimir G. Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 371–383, 1990.

[25] G. N. Watson. *Theory of Bessel functions*. Cambridge University Press, 1952.

[26] Qun Xie and Andrew Barron. Minimax redundancy for the class of memoryless sources. Unpublished Manuscript, 1995.

[27] Kenji Yamanishi. A decision-theoretic extension of stochastic complexity and its applications to learning. Unpublished Manuscript, 1995.

[28] Zhongxin Zhang. *Discrete Noninformative Prioirs*. PhD thesis, Yale University, 1994.

A Proof of Theorem 3

We want to calculate the following integral:

$$\int_0^1 \omega(p) e^{-T D_{\mathrm{KL}}(\hat{\theta}||p)} dp$$

Which we can expand and write as follows:

$$\int_0^1 \frac{1}{\pi\sqrt{p(1-p)}} \exp\left(T\hat{\theta}\log\left(\frac{p}{\hat{\theta}}\right) T(1-\hat{\theta})\log\left(\frac{1-p}{1-\hat{\theta}}\right)\right) dp$$
$$= \frac{1}{\pi\hat{\theta}^{T\hat{\theta}}(1-\hat{\theta})^{T(1-\hat{\theta})}} \int_0^1 p^{T\hat{\theta}-1/2}(1-p)^{T(1-\hat{\theta})-1/2} dp\ .$$

Luckily, the last integral is a well studied quantity, called the Beta function. More specifically, it is $B(T\hat{\theta}+1/2, T(1-\hat{\theta})+1/2)$, which can also be expressed in terms of the Gamma function, $B(x,y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$. [13] Using these relations we get:

$$\int_0^1 \omega(p) e^{-TD_{\mathrm{KL}}(\hat\theta||p)} dp = \frac{B(T\hat\theta + 1/2, T(1-\hat\theta)+1/2)}{\pi \hat\theta^{T\hat\theta}(1-\hat\theta)^{T(1-\hat\theta)}}$$
$$= \frac{\Gamma(T\hat\theta + 1/2)\Gamma(T(1-\hat\theta)+1/2)}{\pi\Gamma(T+1)\ \hat\theta^{T\hat\theta}(1-\hat\theta)^{T(1-\hat\theta)}}$$

Plugging this formula into Equation 20 we get:

$$\begin{aligned}
&-\ln \frac{\Gamma(T\hat\theta + 1/2)\Gamma(T(1-\hat\theta)+1/2)}{\pi\Gamma(T+1)\hat\theta^{T\hat\theta}(1-\hat\theta)^{T(1-\hat\theta)}} \\
&= \ln\Gamma(T+1) + \ln\pi + T\left(\hat\theta\ln(\hat\theta) + (1-\hat\theta)\ln(1-\hat\theta)\right) \\
&\quad -\ln\Gamma(T\hat\theta + 1/2) - \ln\Gamma(T(1-\hat\theta)+1/2)
\end{aligned}$$

The asymptotic expansion of $\ln\Gamma(z)$ for large values of z can be used to give upper and lower bounds on this function for positive values of z (See Equations 6.1.41 and 6.1.42 in [1]).

$$\begin{aligned}
&(z-1/2)\ln z - z + (1/2)\ln(2\pi) + \frac{1}{12z} - \frac{1}{360z^3} \\
&\leq \ln\Gamma(z) \leq \\
&(z-1/2)\ln z - z + (1/2)\ln(2\pi) + \frac{1}{12z}
\end{aligned} \tag{35}$$

Using these bounds we get the statement of the theorem (details in appendix A).

$$\begin{aligned}
&\ln\Gamma(T+1) - \ln\Gamma(T\hat\theta+1/2) - \ln\Gamma(T(1-\hat\theta)+1/2) + \ln\pi + T\left(\hat\theta\ln(\hat\theta) + (1-\hat\theta)\ln(1-\hat\theta)\right) \\
\leq\ & (T+1/2)\ln(T+1) - (T+1) + (1/2)\ln(2\pi) + \frac{1}{12(T+1)} \\
& -T\hat\theta\ln(T\hat\theta+1/2) + (T\hat\theta+1/2) - (1/2)\ln(2\pi) - \frac{1}{12(T\hat\theta+1/2)} + \frac{1}{360(T\hat\theta+1/2)^3} \\
& -T(1-\hat\theta)\ln(T(1-\hat\theta)+1/2) + (T(1-\hat\theta)+1/2) - (1/2)\ln(2\pi) - \frac{1}{12(T(1-\hat\theta)+1/2)} \\
& +\frac{1}{360(T(1-\hat\theta)+1/2)^3} \\
& +\ln\pi + T\left(\hat\theta\ln\hat\theta + (1-\hat\theta)\ln(1-\hat\theta)\right)
\end{aligned}$$

[13] Essentially, the Gamma function is an extension of the Factorial to the reals and the Beta function is an extension of the reciprocal of the Binomial function.

$$\leq \frac{1}{2}\ln\frac{\pi}{2}+\frac{1}{2}\ln(T+1)+T\left(\hat{\theta}\ln\frac{2T\hat{\theta}+2\hat{\theta}}{2T\hat{\theta}+1}+(1-\hat{\theta})\ln\frac{2T(1-\hat{\theta})+2(1-\hat{\theta})}{2T(1-\hat{\theta})+1}\right)+\frac{1}{360(T+1)^3}$$

$$\leq \frac{1}{2}\ln(T+1)+\frac{1}{2}\ln\frac{\pi}{2}+\frac{1}{2}-\frac{1}{2+(T\min(\hat{\theta},1-\hat{\theta}))^{-1}}+\frac{1}{360(T+1)^3} \quad (36)$$

$$\leq \frac{1}{2}\ln(T+1)+1 \quad (37)$$

B Proof of Lemma 1.

Given that the bound on the regret of the min-max optimal algorithm is equal to $\ln Z_T$, where Z_T is the normalization factor of the min/max distribution for sequences of length T, our goal is to calculate

$$\lim_{T\to\infty}\ln Z_T = \lim_{T\to\infty}\ln Z_T\sum_{\mathbf{x}^T}\max_p P_p(\mathbf{x}^T)\ .$$

The probability assigned to a sequence $\mathbf{x}^T$ by the model p depends only on the number of 1's in $\mathbf{x}^T$, i.e., on $T\hat{\theta}$. It is

$$P_p(\mathbf{x}^T) = p^{T\hat{\theta}}(1-p)^{T(1-\hat{\theta})} = \left(p^{\hat{\theta}}(1-p)^{1-\hat{\theta}}\right)^T \quad (38)$$

The value of $P_p(\mathbf{x}^T)$ for a given $\mathbf{x}^T$ is maximized when $p=\hat{\theta}$ thus we get that

$$Z_T = \sum_{\mathbf{x}^T}\left(\hat{\theta}^{\hat{\theta}}(1-\hat{\theta})^{1-\hat{\theta}}\right)^T$$

As there are $\binom{T}{T\hat{\theta}}$ sequences of length T in which the fraction of 1s is $\hat{\theta}$, and as $\hat{\theta}$ achieves the values i/T for $i=0,\ldots,T$, we can rewrite the last equation in the following form.

$$Z_T = \sum_{i=0}^{T}\binom{T}{i}e^{-T\ H(i/T)} \quad (39)$$

where $H(p)=-p\ln p-(1-p)\ln(1-p)$ is the binary entropy of p.

To approximate the value of Equation 39 we replace $\binom{T}{i}$ by the equivalent expression $\Gamma(T+1)/(\Gamma(i+1)\Gamma(T-i+1))$ move the log of this expression to the exponent, and get:

$$Z_T = \sum_{i=0}^{T}\exp\big(\ln\Gamma(T+1)-\ln\Gamma(i+1) \\ -\ln\Gamma(T-i+1)-TH(i/T)\big) \quad (40)$$

Replacing $\Gamma(\cdot)$ by its series expansion around $T = \infty$ and $H(p)$ by its definition, we get, in the exponent, the expression

$$\begin{aligned}
& (T+\frac{1}{2})\ln(T+1)-(T+1)+\frac{1}{2}\ln 2\pi+O(1/T) \\
- \quad & (i+\frac{1}{2})\ln(T+1)+(i+1)-\frac{1}{2}\ln 2\pi+O(1/T) \\
- \quad & (T-i+\frac{1}{2})\ln(T-i+1)+(T-i+1)-\frac{1}{2}\ln 2\pi+O(1/T) \\
+ \quad & i\ln\frac{i}{T}+(T-i)\ln\frac{T-i}{T} \\
= \quad & 1-\frac{1}{2}\ln 2\pi+O(1/T) \\
+ \quad & (T+1/2)\ln(T+1)-(i+1/2)\ln(i+1)-(T-i+1/2)\ln(T-i+1) \\
+ \quad & i\ln i+(T-i)\ln(T-i)-T\ln(T) \\
= \quad & 1-\frac{1}{2}\ln 2\pi+O(1/T) \\
+ \quad & T\ln(1+\frac{1}{T})-i\ln(1+\frac{1}{i})-(T-i)\ln(1+\frac{1}{T-i}) \\
+ \quad & \frac{1}{2}\ln\frac{T+1}{(i+1)(T-i+1)}
\end{aligned}$$

In the first line of the last expression, the first two terms are constant and the third term is $o(1)$. All the terms in the second line are in the range $[0,1]$. The first term is equal to $1-O(1/T)$, and if $\hat{\theta}\in[\epsilon,1-\epsilon]$ for some fixed $\epsilon>0$ then the second and third term have the same asymptotic behavior. The dominant term, in any case, is the last one.

Returning to Equation (39), we separate the sum into three parts as follows

$$Z_T = \sum_{i=0}^{T}\binom{T}{i}e^{-T\ H(i/T)} = \sum_{i=0}^{\epsilon T}\binom{T}{i}e^{-T\ H(i/T)} + \sum_{i=\epsilon T}^{(1-\epsilon)T}\binom{T}{i}e^{-T\ H(i/T)} + \sum_{i=(1-\epsilon)T}^{T}\binom{T}{i}e^{-T\ H(i/T)}$$

Using the approximations shown above we can upper bound the summands in the first and third sums by

$$\binom{T}{i}e^{-T\ H(i/T)} \leq e^{2-\frac{1}{2}\ln 2\pi+O(1/T)}\sqrt{\frac{T+1}{(i+1)(T-i+1)}}$$

and estimate the summands in the second term by

$$\binom{T}{i}e^{-T\ H(i/T)} = e^{-\frac{1}{2}\ln 2\pi+O(1/T)}\sqrt{\frac{T+1}{(i+1)(T-i+1)}}$$

A convenient way for writing the common factor is

$$\sqrt{\frac{T+1}{(i+1)(T-i+1)}} = \sqrt{T+1}\frac{1}{T+2}\sqrt{\frac{1}{\frac{i+1}{T+2}\frac{T-i+1}{T+2}}}$$

Using these equalities we can write the second, and major summation as follows:

$$\sum_{i=\epsilon T}^{(1-\epsilon)T} \binom{T}{i} e^{-T\ H(i/T)} = e^{O(1/T)} \frac{1}{\sqrt{2\pi}} \sqrt{T+1} \sum_{i=\epsilon T}^{(1-\epsilon)T} \frac{1}{T+2} \sqrt{\frac{1}{\frac{i+1}{T+2} \frac{T-i+1}{T+2}}} \ .$$

Observe the last sum, it is easy to see that it a Riemann Sum which is a finite approximation to the integral

$$\int_{\epsilon}^{1-\epsilon} \sqrt{\frac{1}{p(1-p)}} dp \ .$$

And, as $T \rightarrow \infty$, and the function $1/\sqrt{p(1-p)}$ is Riemann integrable this sum approaches the value of the integral, so we get:

$$\sum_{i=\epsilon T}^{(1-\epsilon)T} \binom{T}{i} e^{-T\ H(i/T)} = e^{O(1/T)} \sqrt{T+1} \frac{1}{\sqrt{2\pi}} \int_{\epsilon}^{1-\epsilon} \sqrt{\frac{1}{p(1-p)}} dp \ .$$

Similarly, we get that the sum that corresponds to the first and last terms approach

$$\sum_{i=0}^{\epsilon T} \binom{T}{i} e^{-T\ H(i/T)} \leq e^{2+O(1/T)} \sqrt{T+1} \frac{1}{\sqrt{2\pi}} \int_{0}^{\epsilon} \sqrt{\frac{1}{p(1-p)}} dp \ .$$

and

$$\sum_{i=(1-\epsilon)T}^{T} \binom{T}{i} e^{-T\ H(i/T)} \leq e^{2+O(1/T)} \sqrt{T+1} \frac{1}{\sqrt{2\pi}} \int_{1-\epsilon}^{1} \sqrt{\frac{1}{p(1-p)}} dp \ .$$

The last two integrals are $O(\epsilon)$, thus, after we take the limit $T \rightarrow \infty$ we can take the limit $\epsilon \rightarrow \infty$ and get that

$$\begin{aligned}
\lim_{T\rightarrow\infty} \frac{Z_T}{\sqrt{T+1}} &= \lim_{\epsilon\rightarrow 0} \lim_{T\rightarrow\infty} \left(\sum_{i=0}^{\epsilon T} \binom{T}{i} e^{-T\ H(i/T)} + \sum_{i=\epsilon T}^{(1-\epsilon)T} \binom{T}{i} e^{-T\ H(i/T)} + \sum_{i=(1-\epsilon)T}^{T} \binom{T}{i} e^{-T\ H(i/T)} \right) \\
&= \lim_{\epsilon\rightarrow 0} \left(e^2 \frac{1}{\sqrt{2\pi}} \int_0^{\epsilon} \sqrt{\frac{1}{p(1-p)}} dp + \frac{1}{\sqrt{2\pi}} \int_{\epsilon}^{1-\epsilon} \sqrt{\frac{1}{p(1-p)}} dp + e^2 \frac{1}{\sqrt{2\pi}} \int_{1-\epsilon}^{1} \sqrt{\frac{1}{p(1-p)}} dp \right) \\
&= \frac{1}{\sqrt{2\pi}} \int_0^1 \sqrt{\frac{1}{p(1-p)}} dp \\
&= \sqrt{\frac{\pi}{2}}
\end{aligned}$$

Finally taking the log we get that

$$\lim_{T\rightarrow\infty} \ln Z_T - \frac{1}{2} \ln(T+1) = \frac{1}{2} \ln \frac{\pi}{2}$$

which completes the proof of the theorem.

C Proof of Theorem 4

We need to calculate a lower bound on the following integral:

$$\int_0^1 \exp \quad \Big(2T\hat{\theta}\left((1-\hat{\theta})^2-(1-p)^2\right) \\ +2T(1-\hat{\theta})\left((0-\hat{\theta})^2-(0-p)^2\right)\Big)\,dp\;.$$

This integral simplifies to

$$\int_0^1 \exp\left(-2T(\hat{\theta}-p)^2\right)dp\;.$$

This is another well-known integral, it is (up to a multiplicative constant) an integral of the normal distribution on part of the real line. This integral does not have a closed algebraic form, but can be expressed using the error function $\mathrm{erf}\,(\cdot)$.

$$\begin{aligned} & \int_0^1 \exp\left(-2T(\hat{\theta}-p)^2\right) \qquad (41)\\ = \;& \sqrt{\frac{\pi}{2T}}\left(1-\int_0^\infty e^{-2T(\hat{\theta}+p)^2}dp-\int_0^\infty e^{-2T(1-\hat{\theta}+p)^2}dp\right)\\ = \;& \frac{1}{2}\sqrt{\frac{\pi}{2T}}\left(\mathrm{erf}\left(\hat{\theta}\sqrt{T}\right)+\mathrm{erf}\left((1-\hat{\theta})\sqrt{T}\right)\right)\end{aligned}$$

where

$$\mathrm{erf}\,(p) \doteq \frac{2}{\sqrt{\pi}}\int_0^p e^{-x^2}dx$$

is the cumulative distribution function of the normal distribution. As $\mathrm{erf}\,(x)$ is an increasing function we find that $\mathrm{erf}\left(\hat{\theta}\sqrt{2T}\right)+\mathrm{erf}\left((1-\hat{\theta})\sqrt{2T}\right)$ is minimized when $T=1$, and as $\mathrm{erf}\,(x)$ is concave for $x\geq 0$ we find that the minimum is achieved for $\hat{\theta}=0$ or $\hat{\theta}=1$. Thus we get that the integral is uniformly (w.r.t. $\hat{\theta}$) lower bound by

$$\int_0^1 \exp\left(-2T(\hat{\theta}-p)^2\right) \leq \frac{\mathrm{erf}\left(\sqrt{2}\right)}{2}\sqrt{\frac{\pi}{2T}} \qquad (42)$$

We now plug this lower bound into Equation (20) and get that the regret is uniformly upper bounded by

$$-\frac{1}{2}\ln\left(\frac{\mathrm{erf}\left(\sqrt{2}\right)}{2}\sqrt{\frac{\pi}{2T}}\right) = \frac{1}{4}\ln T+\frac{1}{2}\ln\frac{2}{\mathrm{erf}\left(\sqrt{2}\right)}-\frac{1}{4}\ln\frac{\pi}{2} \qquad (43)$$

D An example for which the exponential weights algorithm is better than any Bayes algorithm

Suppose the model class consists of just three biased coins: $p_1 = 0, p_2 = 1/2, p_3 = 1$, and that a sequence is generated by flipping a random coin whose bias is 0.9. Consider first the Bayes algorithm, whatever is the choice of the prior distribution, once the algorithm observes both a zero and a one in the sequence, the posterior distribution becomes concentrated on $p_2 = 1/2$ and all predictions from there on would necessarily be $\phi_t = 1/2$. This would causes the algorithm an average loss per trial of $(0.5 - 0.1)^2 = 0.16$. On the other hand, consider the EW algorithm which uses the uniform prior distribution over the three models. The total loss of this algorithm is guaranteed to be larger than that of the best model in the class by at most $ln(2)/4$. For large T, with very high probability, the best model is $p_2 = 1$ whose average loss per trial is $(0.9 - 1)^2 = 0.01$. Thus the average loss per trial of the EW algorithm is guaranteed to quickly approach 0.01, making it a clearly better choice than the Bayes algorithm for this problem. We conjecture that a gap between the performance of there algorithms exists when the class of models is the set of all the biased coins. However, we have yet not been able to calculate optimal prior for the optimal Bayes algorithm for this case.