ISIT'2013 1569719891

A Universal Probability Assignment for Prediction of Individual Sequences

Yuval Lomnitz and Meir Feder
Department of EE-Systems, Tel Aviv University
Email: yuval.lomnitz@gmail.com ,meir@eng.tau.ac.il

***Abstract*—Is it a good idea to use the frequency of events in the past, as a guide to their frequency in the future (as we all do anyway)? In this paper the question is attacked from the perspective of universal prediction of individual sequences. It is shown that there is a universal sequential probability assignment, such that for a large class loss functions (optimization goals), the predictor minimizing the expected loss under this probability, is a good universal predictor. The proposed probability assignment is based on randomly dithering the empirical frequencies of states in the past, and it is easy to show that randomization is essential. This yields a very simple universal prediction scheme which is similar to Follow-the-Perturbed-Leader (FPL) and works for a large class of loss functions, as well as a partial justification for using probabilistic assumptions.**

I. Introduction

In this paper the problem of universal sequential prediction of an individual unknown sequence is considered [1][2][3], and a prediction approach based on universal probability assignment is proposed. Given a space of strategies $\mathcal{B}$, a space of nature states $\mathcal{X}$ and a loss function $l(b,x), b \in \mathcal{B}, x \in \mathcal{X}$, the purpose is to assign the next strategy $\hat{b}_t$ given the knowledge of the past states $\mathbf{x}_1^{t-1}$, such that the overall loss $\sum_{t=1}^{n} l(\hat{b}_t, x_t)$ would be asymptotically close to the loss obtained by the best fixed strategy known a-posteriori after viewing the entire sequence $\mathbf{x}_1^n$, i.e. $\min_{b\in\mathcal{B}} \sum_{t=1}^{n} l(b, x_t)$.

In the particular case of sequential probability assignment under the log loss function $l(b,x) = \log \frac{1}{b(x)}$ where $\mathcal{B}$ is the space of probability assignments on the finite alphabet $\mathcal{X}$, or equivalently in universal sequential compression, it is shown [3][4, §13][1, §9] that it is possible to assign probabilities $\hat{p}_t(x_t)$ for the next state in an arbitrary sequence of states $x_t \in \mathcal{X}, t = 1, 2, \ldots, n$, given the past states, such that for any possible sequence, the overall probability $\hat{p}(\mathbf{x}) = \prod_{t=1}^{n} \hat{p}_t(x_t)$ would not be too far, in a multiplicative or logarithmic sense, from the best i.i.d. probability assigned to the sequence a-posteriori $\max_{p(\cdot)} \prod_{t=1}^{n} p(x_t)$. The result extends to probability assigned by Markov machines or finite state machines [5]. This problem is related to universal compression because the overall compression length corresponds to $\log\left(\frac{1}{p(\mathbf{x})}\right)$. A remarkable feature of these universal probability assignments is that, although nothing is assumed about the sequence, to construct a universal encoder it is enough to encode as if $\hat{p}_t(\cdot)$ was the *true* probability of the next state.

These universal probability assignments, such as the Laplace [4, §13.2] or Krichevsky-Trofimov (KT) [6] estimators, have an intuitively appealing structure which induces a small bias over the empirical distribution seen so far. For example, Laplace's estimate for the probability distribution of of x_t is

$$\hat{p}_t(x) = \frac{N_{t-1}(x) + 1}{(t-1) + |\mathcal{X}|}, \tag{1}$$

where $N_t(x)$ denotes the number of times the state x appears in $\mathbf{x}_1^t$. While these estimators get closer with time to the measured empirical distribution, they do not "trust" it completely, and, for example, never assign a probability value 0 to states that had not appeared before. Furthermore, in the probabilistic prediction setting the same distributions were shown to perform well not only for the log loss: the predictor which minimizes the expected loss under these distributions $\hat{b}_t = \operatorname{argmin}_b \mathbb{E}_{X\sim\hat{p}_t(\cdot)} l(b, X)$ operates well for a wider class of loss functions [3, §III.A.2].

This naturally leads to the following question: is it possible to forecast an individual sequence by first generating a probability assignment based on the past, and then minimizing the expected loss under this assignment (i.e. in a way, acting as if future events truly happen with this probability)? Consider prediction schemes of the following form:

1) Generate a probability assignment $P_t^{(u)}(x)$ based on the past of the sequence $\mathbf{x}_1^{t-1}$, in a way which does not depend on the loss function.
2) To predict b_t under the loss function $l(b,x)$, choose the strategy that minimizes the expected loss under $P_t^{(u)}$, i.e.:

$$\hat{b}_t = \operatorname*{argmin}_b \mathbb{E}_{X\sim P_t^{(u)}(\cdot)} [l(b, X)] \tag{2}$$

If there exists a single scheme for generating $P_t^{(u)}(\cdot)$ that does not depend on the loss function $l(b,x)$, but for which $\hat{b}_t$ yields a good (Hannan-consistent [1]) predictor for a certain class of loss functions, then we call $P_t^{(u)}(\cdot)$ a *universal sequential probability assignment* with regards to that class. Notice that this term had been used in the past with respect to the log-loss, so the definition above can be considered a natural extension.

It is easy to show that, if the class of loss functions includes even simple loss functions such as the 0-1 loss (the number of errors), then no deterministic assignment can be universal, and therefore the Laplace or KT assignments are inadequate. However, it is shown in this paper that the random assignment obtained by slightly perturbing the empirical frequencies is universal for a large class of loss functions, including the log-loss and any bounded loss.

In addition to supplying a simple and general universal prediction scheme, this result also has interpretations contributing to our understanding of probability. For example, it supplies justification for treating the statistics of a process in the past as a guide to its statistics in the future, without having to assume the process is indeed stationary, or that it is driven by a "probabilistic" law. In other words, if our natural behavior is in some way similar to the prediction algorithm described here, then the claims on its convergence can be used to justify this behavior.

The next section completes the problem definition and discusses the boundaries of the solution, and relations to known results. Section III gives the main results (the proofs are omitted and can be found in the full paper [7]), and Section IV discusses the possible implications on understanding probabilistic behavior.

II. Problem statement and discussion

Building upon the definitions already presented in the introduction, in this section some complementary definitions are presented. We assume throughout this paper that $\mathcal{X}$ is finite (otherwise there is no meaning to measuring empirical frequencies). The set of possible strategies $\mathcal{B}$ is not restricted. The loss function $l(b,x)$ is constant over time.

Let us define the accumulated loss of a sequential predictor $\hat{b}_t(\mathbf{x}_1^{t-1})$ as:

$$\hat{L}_n = \sum_{t=1}^{n} l(\hat{b}_t, x_t), \tag{3}$$

and the loss of the best fixed strategy as:

$$L_n^* = \min_b \sum_{t=1}^{n} l(b, x_t). \tag{4}$$

The difference $\hat{L}_n - L_n^*$ which is defined as the regret, is a function of the predictor and the sequence. The worst case regret is:

$$\mathcal{R}_{\max} = \max_{\mathbf{x}_1^n} \left(\hat{L}_n - L_n^* \right), \tag{5}$$

and the normalized regret is $\frac{\mathcal{R}_{\max}}{n}$. A forecasting strategy $\hat{b}_t$ is said to be *Hannan-consistent*, if $\limsup_{n\to\infty} \frac{\mathcal{R}_{\max}}{n} \leq 0$ almost surely (the probability is over the randomization in the forecaster if it is random). This means that for large n, the loss of the forecaster is essentially at least as small as that of any fixed strategy. As mentioned in the introduction, the problem addressed in this paper is of finding a sequential probability assignment $P_t^{(u)}(\cdot)$ such that the resulting prediction scheme (2) is Hannan-consistent for a large class of loss functions. We will focus mainly on bounding the *expected* loss (over the predictor's randomization), because it also leads to almost-sure bounds by applying the strong law of large numbers. The maximum *expected* regret is defined as:

$$\overline{\mathcal{R}}_{\max} = \max_{\mathbf{x}_1^n} \mathbb{E}\left[\hat{L}_n - L_n^* \right], \tag{6}$$

For some loss functions satisfying smoothness conditions [1, Thm 3.1][2, Thm 1], the forecasting strategy known as "Follow the Leader" (FL), which chooses at each time the best strategy in retrospect $\hat{b}_t^{(\mathrm{FL})} = \operatorname{argmin}_b \sum_{i=1}^{t-1} l(b, x_i)$, is Hannan consistent. Rewriting the above as $\hat{b}_t^{(\mathrm{FL})} = \operatorname{argmin}_b \sum_{x\in\mathcal{X}} \frac{N_{t-1}(x)}{t-1} l(b,x)$, it can be interpreted as an implementation of (2) where the universal probability assignment equals the empirical frequencies $P_t^{(u)}(\cdot) = \frac{N_{t-1}(x)}{t-1}$. In other words, for this family of loss functions, there is a simple solution for $P_t^{(u)}(\cdot)$, namely the empirical distribution. However this class of loss functions where FL is universal, is rather limited.

For a probability assignment to be "general" enough, one would want to cover, at the least, the family of discrete-strategy, discrete-state loss functions, presented by Hannan [8]. For this family, the loss function can be represented by a general $|\mathcal{B}| \times |\mathcal{X}|$ matrix specifying the loss for each strategy and each state of nature. It is well known [1, §4] and straightforward to see that randomization is required in order to cover this class: consider the 0-1 loss case, i.e. binary sequences $\mathcal{X} = \mathcal{B} = \{0,1\}$ with $l(b,x) = \mathrm{Ind}(b \neq x)$, where the total loss is the number of errors. For this loss function, no deterministic predictor yields Hannan-consistency, because for each deterministic predictor there exists a sequence which fails the predictor completely, by choosing the next outcome as the opposite of the predictor's choice, while the loss of the best fixed predictor is at most $n/2$. Because a deterministic $P_t^{(u)}(\cdot)$ inevitably leads to a deterministic predictor (2), this implies a random $P_t^{(u)}(\cdot)$ is required, in general.

For the binary 0-1 loss problem, Feder, Merhav and Gutman [9] used a small dither when the empirical probability is close to $\frac{1}{2}$, which effectively avoids a decision when the frequencies of $0, 1$ are nearly equal.[1] For this specific problem, the optimal solution (in the sense of minimax regret) is known exactly and was presented by Cover [10]. While the optimal dither in this problem is different than the straight line used by Feder, Merhav and Gutman, and is not known in general, this is of no consequence in the current problem, as we are only considering Hannan consistency. This solution, as well as the small bias from the empirical distribution which is required in the log-loss problem (1), motivates the following choice of $P_t^{(u)}(\cdot)$: add a small dither to $N_{t-1}(x)$ (the counts of events in the past) and re-normalize. As shown below, this solution achieves Hannan-consistency for any bounded loss function and for the log loss.

The proposed forecaster is reminiscent of the scheme termed "Follow the Perturbed Leader" (FPL), originally proposed by Hannan [8], in which the decision is obtained by adding a small dither to the accumulated loss of every reference strategy and then choosing the best one. Indeed, dithering the frequencies is similar, but not equivalent, to dithering the accumulated losses, and our proof technique for the bounded loss case borrows from Kalai and Vempala's [11]. Following

[1] It is interesting to note that for the 0-1 loss problem their forecaster is equivalent to a "Follow the Perturbed Leader" forecaster with a uniform distribution (see below) and also equivalent to the forecaster proposed here.

this similarity we term the scheme proposed here "Follow the Perturbed Frequency" (FPF). Notice, however, that FPL is defined, in general, only when the number of strategies is finite, while FPF is defined, in general, only when the number of outcomes (states) is finite, and does not have to assume the number of strategies is finite. On the other hand, FPL can deal with more general forms of the problem, including time-varying loss functions.

The problem considered here is a close relative of the calibration problem [1, §4.5], i.e. the problem of estimating from an individual sequence, probability forecasts that pass certain consistency tests. The problems are related in that, in both cases it is shown possible to generate from empirical data collected from an individual sequence, probability assignments that appear to operate as well as forecasts which are based on knowledge of the "true" statistical model. Also, randomization is essential in both cases. However, none of the problems is a special case of the other: the probability assignment shown here is not necessarily calibrated, and a calibrated probability assignment does not necessarily satisfy the requirements of the current problem.[2]

In this paper, in order to simplify matters, only fixed strategies are considered as reference. As one of our motivations is to rationalize the behavior of learning probabilities from the past, it is enough to consider fixed strategies in order to see the advantage of this behavior. The extension to dynamic reference strategies is unfortunately not immediate as in the setting of prediction with expert advice [1, §2], where dynamic strategies can be turned into fixed ones by simple enumeration (i.e. replacing the strategy with the index of the strategy), because we explicitly assume a fixed loss function. However in some cases, the core of the prediction problem lies in competing with fixed strategies. For example, reference strategies defined by states (such as Markov predictors or finite state machines), can be considered as fixed strategies in each sub-sequence belonging to the same state.

III. Main results

Let $N_t(x)$ be number of times a specific x occurred in the sequence $\mathbf{x}$ up to and including time t. The universal sequential probability assignment is defined as:

$$\begin{aligned} P_t^{(u)}(x) &= c_t \cdot (N_{t-1}(x) + h_t \cdot u_t(x)) \\ &= \frac{N_{t-1}(x) + h_t \cdot u_t(x)}{t - 1 + h_t \cdot \sum_{x' \in \mathcal{X}} u_t(x')} \end{aligned} \tag{7}$$

where $c_t^{-1} = \sum_{x \in \mathcal{X}} (N_{t-1}(x) + h_t \cdot u_t(x))$ is the normalizer guaranteeing unit sum. $u_t(x) \sim U[0,1]$ is a random dither which is assumed to be uniformly distributed, i.i.d. over different x and t (dependence over t does not affect the expected regret). h_t is a non-decreasing positive sequence. Our philosophical considerations (i.e. justifying probabilistic behavior) motivate keeping h_t as general as possible rather than finding a specific optimal sequence h_t for each problem.

The FPF predictor, for any loss function $l(b,x)$, is defined by:

$$b_t^{(\mathrm{FPF})} = \operatorname*{argmin}_{b \in \mathcal{B}} \mathop{\mathbb{E}}_{\mathbf{X} \sim P_t^{(u)}(x)} [l(b, X)] \tag{8}$$

Theorem 1. *Assuming $h_t = h_1 \cdot t^{\alpha}$, with $\alpha \in (0,1)$, the FPF predictor is Hannan-consistent for any bounded loss function and for the log-loss. Therefore under these conditions, $P_t^{(u)}(x)$ defined in* (7) *is a universal probability assignment for the class.*

This theorem is based on the two following theorems:

Theorem 2. *Assume the loss function is bounded $|l(b,x)| \leq R$. Then:*

1) *The expected regret of FPF is upper bounded by*

$$\overline{\mathcal{R}}_{\max} \leq 2R \sum_{t=1}^{n} h_t^{-1} + 2R|\mathcal{X}|h_n \tag{9}$$

2) *Particularly, for any $h_t = h_1 \cdot t^{\alpha}$, with $\alpha \in (0,1)$, the normalized expected regret $\frac{1}{n}\overline{\mathcal{R}}_{\max}$ tends to zero with n.*
3) *For $h_t = \sqrt{\frac{2t}{|\mathcal{X}|}}$, $\frac{1}{n}\overline{\mathcal{R}}_{\max} \leq 4R\sqrt{\frac{2|\mathcal{X}|}{n}}$.*

Corollary 2.1. *The theorem holds under a milder condition, that the loss function is bounded only for the set of optimizing strategies, defined as*

$$\mathcal{B}_{\mathrm{opt}} = \left\{ \operatorname*{argmin}_{b \in \mathcal{B}} \sum_{x \in \mathcal{X}} \lambda(x) l(b,x) : \lambda(x) \geq 0, \exists x : \lambda(x) > 0 \right\} \tag{10}$$

and where $R = \sup_{x \in \mathcal{X}, b \in \mathcal{B}_{\mathrm{opt}}} l(b,x)$. Particularly, the theorem holds for the L_2 norm loss, $l(\mathbf{b},\mathbf{x}) = \|\mathbf{b} - \mathbf{x}\|^2$ for $\mathcal{X} \subset \mathbb{R}^d$ ($|\mathcal{X}| < \infty$), and $\mathcal{B} = \mathbb{R}^d$. In that case $R = \max_{\mathbf{x},\mathbf{x}' \in \mathcal{X}} \|\mathbf{x} - \mathbf{x}'\|^2$ is the squared diameter of the set $\mathcal{X}$.

Notice that in the most general case without any limitations (such as on magnitude), it is generally impossible to devise a universal scheme for the L_2 norm loss that beats the best fixed strategy, i.e. the empirical mean, up to a constant, and it is made possible in the current problem by the assumption that $\mathcal{X}$ is finite.

Proof concept: The proof of Theorem 2 is similar in spirit to the proof of Kalai and Vempala [11] for the FPL forecaster, as the perturbation on $N_t(x)$ can be translated to a perturbation on the accumulated loss. The idea behind the proof is generally as follows: it is well known [1][2][11] that the clairvoyant predictor having access to the current state x_t and choosing the strategy b which minimizes the loss up to and *including* time t, is at least as good as the best fixed strategy. Like FL (see previous section), this predictor can be interpreted as minimizing the expected loss according to the empirical number of states including time t, $N_t(x)$. If one adds dither, and calculates the loss as if the number of times x

[2]Consider for example the 0-1 loss problem, and a sequence containing an equal number of zeros and ones. Any probability forecaster yielding only values in the range $0.5 \pm \epsilon$ is ϵ-calibrated, while the decisions based on these probabilities (when plugged into (2)) can be arbitrary (depending on whether the probability is smaller or larger than 0.5), and can yield arbitrarily bad (or good) aggregate losses.

happened had been $N_t(x)+h_t\cdot u_t(x)$, there is a small penalty, which is quantified. Now, when taking back the superfluous information x_t and following the counts $N_{t-1}(x)+h_t\cdot u_t(x)$, which is what the FPF predictor does, the loss in the transition can be bounded. In this step, the dither is crucial, and for the loss to be small, the dither $h_t u_t(x_t)$ in the direction of the "missing state" x_t should be much larger than the contribution of x_t itself.

Theorem 3. *For the case of the log-loss, where $b(x)$ is a probability distribution over $\mathcal{X}$ and $l(b,x)=\log\left(\frac{1}{b(x)}\right)$,[3] the expected regret of FPF satisfies:*

1)
$$\overline{\mathcal{R}}_{\max}\leq\sum_{t=1}^{n}\frac{|\mathcal{X}|h_t-1}{t}+\sum_{t=1}^{n}\frac{1}{\left\lfloor\frac{t-1}{|\mathcal{X}|}\right\rfloor+h_t} \qquad (11)$$

2) *Particularly, for any $h_t=h_1\cdot t^{\alpha}$, with $\alpha\in[0,1)$, the normalized expected regret $\frac{1}{n}\overline{\mathcal{R}}_{\max}$ tends to zero with n.*
3) *For constant h_t the expected regret behaves like $O(\log n)$ and specifically for the choice $h=|\mathcal{X}|^{-1}$, $\overline{\mathcal{R}}_{\max}\leq|\mathcal{X}|\log(n)$.*

Proof concept: The proof is similar in spirit to the proof of the performance of Laplace's estimator [1]. In this problem, random dither is not necessary, and the main point is to show that random dither is on average similar to a deterministic bias, as in (1), or in other words, that the small probability that the dither would be near-zero and ineffective, does not lead to unbounded losses. The proof also has many similarities to the proof of Theorem 2.

Regarding the last case, notice that this redundancy is similar to the redundancy obtained with Laplace's estimator and approximately twice the redundancy obtained using Kritchevsky-Trofimov's (which is approximately $\frac{|\mathcal{X}|-1}{2}\log n$). However notice that the target of the FPF forecaster was not to produce optimal redundancy for specific loss functions.

The proofs of the theorems stated above appear in the full paper [7].

IV. Implications on the understanding of probability

A. Initial probabilities

A basic question in the application and philosophy of probability theory is: where do initial probabilities originate from (see, e.g. [12]) ? The fact is, that in many situations a probability distribution is deduced from the relative frequency of events in the past. While this deduction may be justified based on some stationarity assumption, it is often used exactly in those situations where precise analysis of the source of events is not possible, and therefore the assumption that the frequency of events in the future would be similar to their frequency in the past is not necessarily justified. In spite of this, we often deduce a probability distribution based on past statistics and use this probability for decision making with regards to future events. It seems that not only humans but also animals use this principle [13].

One motivation for the problem posed in Section II, of searching for a universal probability assignment, is the attempt to justify this behavior based on mathematical, rather than physical assumptions. The theory of universal prediction of individual sequences, or repetitive games, seems a good framework for this purpose, because it facilitates deduction from the past, without assumptions that the past indicates anything with respect to the future. The existing universal prediction schemes are less suitable for this purpose since they determine the next strategy in a contrived way, as a function of the past frequencies and the loss function, whereas in the probability-based decision making, it is assumed that there exist a single "true" probability.

The success of the FPF predictor for a large set of loss functions, indicates that indeed it is useful to rely on past frequencies, and draw from them a "probability" distribution, even if the future is arbitrary. The dither may be interpreted as the assumption that the future would be similar but not identical to the past, and prevents using a too "decisive" strategy (such as choosing '0' or '1' in the 0-1 loss case), based on a small change in the frequencies. It would be farfetched to claim that this is *the* justification for using probabilities: clearly the reason is related to the regularity that many natural processes exhibit; however it supplements our intuitive understanding by showing that even if these assumptions fail, there is still benefit in learning probabilities from the past.

B. Meaning of probability

In the previous section we tried to justify a specific choice of a probability. However, probability itself is not a well defined concept, and many attempts to explain or justify its use have been made. A good introduction to these philosophical questions can be found in [12] (for a quick overview see [14, Chap. 18]). While there is no dispute on the mathematical axiomatic theory dealing with probability functions, the meaning of probability, and the justification for using it are questionable.

In a nutshell, the main interpretations to probability are the relative frequency approach, the a-priori or logical approach and the subjectivistic approach. Relative frequency theories interpret probability as the limiting frequencies in very large groups of events (called "collectives"). A-priori theories interpret probability as logical relation between sentences, and an extension of formal logic: the attributes "true" and "false" are represented by probabilities of 1 and 0, and are extended by adding a range of probabilities in between. Subjectivistic theories interpret probability as a measure of the degree of belief of a certain person in a certain proposition, and therefore its value is not unique.

A main issue in all interpretations is what probability means with respect to the future. The current results can be interpreted under the framework of the subjectivistic theories,

[3] All log-s in this paper are in the natural base.

which view probability as a tool for decision making, i.e. probability is just the relative weight that we put on each future event when making decisions. Because under subjectivistic theories any probability is valid, there is a problem of justifying any specific choice of a probability assignment, as well as the merit of making decisions according to probabilistic considerations.

The current results can be thought of as a partial resolution to this question: the suggestion of learning a probability from the past by biasing or dithering past frequencies, is a good one in the sense that it is better than any fixed behavior (and as a result, of making decisions according to any fixed probability). This demonstrates a clear merit in following probabilistic considerations, which is not dependent on any assumptions with respect to the real world (the process $\mathbf{x}_t$).

There are some issues, however, with this interpretation. First, the problem setting is limited, compared to our actual use of probability. Learning from experience extends far beyond the framework of repetitive games and constant loss functions, as we usually deduce probabilities from the past and use them to solve *new* problems. Also the fact $\mathcal{X}$ is assumed discrete is somewhat limiting, although it may be sufficient to justify probabilistic intuition, which is fundamentally based on distributions on finite sets (such as coins and dice).

But the main weakness of this interpretation is that it relies on randomness for generating the universal probability assignment $P^{(u)}$ (and as a result, the claims we can make are also probabilistic), and so it may lead to a cyclic argument of explaining probability by using probability. The randomness used here is in a restricted form of "controlled randomness" which is generated by the forecaster. I.e. if we believe it is possible to draw random coins, it is enough for this interpretation to hold and be meaningful. An alternative assumption is pseudo-randomness, i.e. assume that we can generate the dither not randomly, but such that "nature" (drawing the next x_t) cannot guess it, and it appears effectively random. Unfortunately, like in many other theories, we are not able to escape some form of "belief" or conjecture with respect to the future.

Another way to avoid the need for randomness is to avoid problems such as the 0-1 loss case, in which one is forced to bet, problems that are insolvable without randomness. For example, if the loss is convex with respect to the strategy, then the loss when taking the expected value of a random strategy b is always better than the expected loss when b is random. In this case, the forecaster can make a deterministic decision: replace (8) with $\hat{b}_t = \mathbb{E}\left\{\hat{b}_t^{(\mathrm{FPF})}\right\}$, where the expected value is with respect to the randomness of $P^{(u)}$. This can be thought of as a different rule for making decisions based on the past: take as probability the empirical frequencies in the past, however when making a decision which changes significantly with respect to small variations in the probability, take the average decision over these small variations. This rule is deterministic and aligns with intuition, however the restriction to "smooth" loss functions may be too limiting.

Another question that would naturally arise with respect to this explanation is how it aligns with the fact that, at least in the theoretical application of probability theory (e.g. estimation theory, communication theory) we do not use dithers in our probabilities. It seems that the idea of dithering the probabilities is similar to checking the sensitivity of a given solution to the probabilistic assumptions. In case the solution to a given problem does not depend crucially on the exact probability values, adding the dither is indeed redundant. On the other hand, if the solution depends crucially on a small change in the probabilistic assumptions, it may be reasonable to doubt its operation in the real world.

REFERENCES

[1] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning and games*. Cambridge University Press, 2006.

[2] N. Merhav and M. Feder, "Universal schemes for sequential decision from individual data sequences," *IEEE Trans. Information Theory*, vol. 39, no. 4, pp. 1280 –1292, Jul. 1993.

[3] ——, "Universal prediction," *IEEE Trans. Information Theory*, vol. 44, no. 6, pp. 2124–2147, Oct. 1998.

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & sons, 1991.

[5] M. Feder, "Gambling using a finite state machine," *IEEE Trans. Information Theory*, vol. 37, no. 5, pp. 1459 –1465, sep 1991.

[6] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Information Theory*, vol. 27, no. 2, pp. 199–207, Mar. 1981.

[7] Y. Lomnitz and M. Feder. (2013, Jan.) A universal probability assignment for prediction of individual sequences. arXiv:1301.6408 [cs.IT].

[8] J. Hannan, "Approximation to bayes risk in repeated play," *Princeton University Press*, vol. Contributions to the Theory of Games, III, Ann. Math. Study Number 39, pp. 97–139, 1957.

[9] M. Feder, N. Merhav, and M. Gutman, "Universal prediction of individual sequences," *IEEE Trans. Information Theory*, vol. 38, no. 4, Jul. 1992.

[10] T. M. Cover, "Behavior of sequential predictors of binary sequences," in *in Proc. 4th Prague Con. Inform. Theory, Statistical Decision Functions, Random Processes*, 1965, pp. 263–272.

[11] A. T. Kalai and S. Vempala, "Efficient algorithms for online decision problems." *Journal of Computer and System Sciences*, vol. 71, no. 3, pp. 291–307, Oct. 2005.

[12] R. Weatherford, *Philosophical foundations of probability theory*. London : Routledge & Kegan Paul, 1982.

[13] Z. Reznikova and B. Ryabko, "Ants and bits," *IEEE Information Theory Society Newsletter*, vol. 62, no. 5, pp. 17–20, 2012.

[14] Y. Lomnitz, "Universal communication over unknown channels," Ph.D. dissertation, Tel Aviv University, Aug. 2012, available online http://www.eng.tau.ac.il/~yuvall/publications/YuvalL_Phd_report.pdf.