Universal Prediction of Individual Binary Sequences in the Presence of Arbitrarily Varying, Memoryless Additive Noise[1]

Tsachy Weissman
tsachy@tx.technion.ac.il

Neri Merhav
merhav@ee.technion.ac.il

Abstract — **The problem of predicting the next outcome of an individual binary sequence, based on past observations which are corrupted by arbitrarily varying memoryless additive noise, is considered. The goal of the predictor is to perform, for each individual sequence, "almost" as well as the best in a set of experts, where performance is evaluated using a general loss function. This setting is a generalization of the original problem of universal prediction of individual sequences relative to a set of experts (cf., e.g., [2] and the many references therein).**

I. Introduction

The noise model considered in this work is that where the observation available to the predictor to make its prediction for time t is the vector $(y_1, \dots, y_{t-1})$, where $y_i = x_i + r_i$, x_i is the clean bit at time i, and $r = \{r_t,\ t \geq 1\}$ is some arbitrarily varying memoryless noise process. The additive noise model considered in this work differs from the binary-valued noise model considered in [1], [3]-[5] (where the observed bit is the bitwise XOR of the clean bit and the noise bit) and joins it to give a more complete picture for the noisy setting [6]. It is shown that even in this noisy environment, when the information available regarding the past sequence is incomplete, a predictor can be guaranteed to successfully compete with a whole set of prediction schemes in considerably strong senses. Furthermore, these performance guarantees are valid for a very large family of noise processes, though the predictor itself does not depend on the statistical characterization of the particular active noise process within this class. In other words, it is *twofold universal* where, in this context, twofold universality stands for universality in the usual sense (w.r.t. the experts in the class and all possible sequences) *and* w.r.t. a family of noise distributions.

II. Statement of the Problem and Main Results

Let $L : \{0,1\} \times [0,1] \to [0,\infty]$ be a fixed *loss function*. A *predictor* (or an *expert*) $F = \{F_t\}_{t\geq 1}$ is a sequence of functions where $F_t : \mathbf{R}^{t-1} \to [0,1]$. We define the cumulative loss of the predictor F, fed by $y^n = (y_1, \dots, y_n)$ and judged with respect to $x^n = (x_1, \dots, x_n) \in \{0,1\}^n$ by $L_F(y^n, x^n) \stackrel{\text{def}}{=} \sum_{t=1}^n L(x_t, F_t(y^{t-1}))$. We consider the case where the noisy observation accessible to the predictor, $y = (y_1, y_2, \dots) \in \mathbf{R}^\infty$ is given by $y_t = x_t + r_t$, $t \geq 1$, where $r = \{r_t, t \geq 1\}$ is a zero-mean, memoryless, arbitrarily varying process: for every n, the p.d.f. governing $r^n = (r_1, \dots, r_n)$ is of the form: $f(r^n|s^n) = \prod_{i=1}^n f(r_i|s_i)$, where $s^n \in \mathcal{S}^n$ is some unknown *arbitrary* sequence of states, and $\mathcal{S}$, is some abstract state-space such that for all $\sigma \in \mathcal{S}$ we have $\int_{-\infty}^{\infty} r \cdot f(r|\sigma) dr = 0$.

Letting $L_F(x^n) \stackrel{\text{def}}{=} EL_F(y^n, x^n)$ denote the expected loss of F when the underlying individual sequence is x^n, we define the *worst-case relative expected loss* of a predictor P by $R_n(P, \mathcal{F}) \stackrel{\text{def}}{=} \max_{x^n \in \{0,1\}^n}(L_P(x^n) - \inf_{F\in\mathcal{F}} L_F(x^n))$. It is shown that, for a large class of loss functions, for any finite set of experts $\mathcal{F}$, there exists a predictor P such that $R_n(P, \mathcal{F}) = O((\ln n)^2 \cdot \ln|\mathcal{F}|)$, while for another class of loss functions we have $R_n(P, \mathcal{F}) = O(\sqrt{n \ln|\mathcal{F}|})$.

Further results show, however, that the prediction strategies that we suggest are guaranteed to be doing well in considerably stronger senses. It is shown that under some mild additional conditions on the noise process, the predictor P satisfies

$$\limsup_{n\to\infty} \frac{L_P(y^n, x^n) - \inf_{F\in\mathcal{F}} L_F(y^n, x^n)}{\sqrt{n \log\log n}} \leq c \quad \text{a.s.} \quad \forall x \in \{0,1\}^\infty,$$

for some deterministic constant c. It is further shown that, using this same predictor, we also have

$$\begin{aligned} &\max_{x^n \in \{0,1\}^n} \Pr\{\frac{1}{n}[L_P(y^n, x^n) - \min_{\mathcal{F}} L_F(y^n, x^n)] > \epsilon\} \\ \leq\ &\exp\{-n(I(\epsilon, B) + o(n))\}, \end{aligned}$$

where, $I(\epsilon, B) > 0$, which lower bounds the possible exponential rate of the decay, is independent of the expert class $\mathcal{F}$.

The remarkable feature of the predictors that we employ is the strong sense in which they are twofold universal. The above described performance bounds hold with the *same* universal predictor P, regardless of the particular state sequence driving the noise process.

References

[1] A. Baruch, "Universal Algorithms for Sequential Decision in the Presence of Noisy Observations," submitted to the senate of the technion (Master's thesis), February 1999. (See also Proc. ISIT 98', p. 331, Cambridge, MA, August 1998.)

[2] D. Haussler, J. Kivinen, and M.K. Warmuth, "Sequential Prediction of Individual Sequences Under General Loss Functions," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1906-1925, September 1998.

[3] T. Weissman and N. Merhav, "On Prediction in the Presence of Noise," unpublished manuscript, 1999 (preprint available).

[4] T. Weissman and N. Merhav, "On Prediction of Individual Sequences Relative to a Set of Experts in the Presence of Noise," in *Proc. 12th Annu. Workshop on Computational Learning Theory*, pp. 19-28, New York: ACM, 1999.

[5] T. Weissman, A. Baruch and N. Merhav, "Twofold Universal Prediction Schemes for Achieving the Finite-State Predictability of a Noisy Individual Binary Sequence," To be submitted to *IEEE Trans. Inform. Theory* (preprint available).

[6] T. Weissman and N. Merhav, "Universal Prediction of Individual Binary Sequences in the Presence of Noise," Submitted to *IEEE Trans. Inform. Theory*, Nov. 1999.

[1] The research, which is supported by the Israeli Science Foundation, is part of the D.Sc. dissertation of the first author. Both authors are with the Department of Electrical Engineering, Technion-Israel Institute of Technology, Haifa 32000, Israel.