Noname manuscript No.
(will be inserted by the editor)

On the Need for Structure Modelling in Sequence Prediction

Supplementary Material

Niall Twomey · **Tom Diethe** · **Peter Flach**

the date of receipt and acceptance should be inserted later

Abstract This document contains the supplementary material for the main paper. Here, we restate the Theorem 3 and present its proof.

1 Auto-correlation of the Occasionally Dishonest Casino

Markov's theorem tells us that a Markov chain is ergodic if there is a strictly positive probability to pass from any state to any other state in one step, so by construction the Occasionally Dishonest Casino (ODC) satisfies ergodicity. Furthermore, by construction the ODC is also a stationary system. Hence the ODC is jointly stationary and ergodic. Following on from this, we can quantify the expected auto-correlation for the ODC as follows.

Theorem 1 *An Occasionally Dishonest Casino (ODC) uses two kinds of die. Define the set of outcomes S,* e.g. *for a 6-sided die $S = \{1,2,3,4,5,6\}$. A fair die has $\frac{1}{|S|}$ probability of rolling any number, and a loaded die that has p_v probability to roll a value $v \in S$ and $p_{\sim v} = \frac{1-p_v}{|S|-1}$ probability to roll each of the remaining numbers. We will use the notation $\Sigma_S = \sum_{x\in S} x$ to denote the sum of the possible outcomes in S, and $\Sigma_{S\setminus v} = \sum_{x\in S\setminus\{v\}} x$. Assume a symmetric probability p_s that the casino switches from fair to loaded die and back. The expected auto-correlation R of the discrete time process depending on v, p_v and p_s, and for a six-sided die $S = \{1,2,3,4,5,6\}$, is given by:*

$$R(v, p_v, p_s) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \left[\frac{2\,(X_t - C)\,(X_{t+n} - C)}{\frac{35}{12} + p_v(v-D)^2 + \frac{p_{\sim v}}{5} \sum_{x\in S\setminus\{v\}} (x-D)^2} \right].$$

$$\textit{where} \quad C = \frac{7-2D}{4}, \quad D = vp_v - \frac{21-v}{5} p_{\sim v} \tag{1}$$

Proof We are seeking an expression for the expected auto-correlation of a discrete time process $R(\tau), \tau \in \mathbb{Z}$ between any two points X_t and $X_{t+\tau}$ in the sequence depending on p_v

Intelligent Systems Laboratory, University of Bristol
E-mail: firstname.lastname@bristol.ac.uk

and p_s. Since the ODC is stationary up to order 2 [Priestley(1988)], the auto-correlation is given by:

$$R(\tau) = R(-\tau) = \mathbb{E}\left[\frac{(X_t-\mu)(X_{t+\tau}-\mu)}{\sigma_2}\right]$$

The mean μ depends on the value v of the loaded face, but does not depend on the switching probability p_s, since it is symmetric and hence in the long run there is an equal chance of being in either the biased or non-biased state. Using $\mu_b(v)$ and μ_u to represent the mean of the biased and unbiased die respectively (the first of which depends on v), we have

$$\begin{aligned}\mu(v) &= \frac{1}{2}\mu_u + \frac{1}{2}\mu_b(v)\\ &= \frac{1}{2}\frac{\Sigma_S}{|S|} + \frac{1}{2}\left(vp_v + \frac{\Sigma_{S\backslash v}}{|S|-1}p_{\sim v}\right).\end{aligned}$$

The variance of the unbiased die σ_u^2 is:

$$\sigma_u^2 = \frac{1}{|S|}\sum_{x\in S}(x-\mu_u)^2.$$

The variance of the biased die (again depending on v) $\sigma_b^2(v)$ is given by:

$$\begin{aligned}\sigma_b^2(v) &= p_v(v-\mu_b(v))^2 + \sum_{x\in S\backslash\{v\}}\frac{p_{\sim v}}{|S|-1}(x-\mu_b(v))^2\\ &= p_v(v - vp_v - \frac{\Sigma_{S\backslash v}}{|S|-1}p_{\sim v})^2 + \sum_{x\in S\backslash\{v\}}\frac{p_{\sim v}}{|S|-1}(x - vp_v - \frac{\Sigma_{S\backslash v}}{|S|-1}p_{\sim v})^2.\end{aligned}$$

Therefore the overall variance $\sigma^2(v)$ is

$$\begin{aligned}\sigma^2 &= \frac{1}{2}\left(\sigma_b^2(v)+\sigma_u^2\right)\\ &= \frac{1}{2}\frac{1}{|S|}\sum_{x\in S}(x-\mu_u)^2 + \frac{1}{2}p_v(v - vp_v - \frac{\Sigma_{S\backslash v}}{|S|-1}p_{\sim v})^2\\ &+ \frac{1}{2}\sum_{x\in S\backslash\{v\}}\frac{p_{\sim v}}{|S|-1}(x - vp_v - \frac{\Sigma_{S\backslash v}}{|S|-1}p_{\sim v})^2.\end{aligned}$$

Defining

$$C_S = \frac{1}{2}\frac{\Sigma_S}{|S|} - \frac{1}{2}(D_S), \quad D_S = vp_v + \frac{\Sigma_{S\backslash v}}{|S|-1}p_{\sim v},$$

and combining this into a single expression for R, and using the ergodicity of the ODC, we have,

$$R(v,p_v,p_s) = \mathbb{E}\left[\frac{2(X_t-C_S)(X_{t+\tau}-C_S)}{\frac{1}{|S|}\sum_{x\in S}(x-\mu_u)^2 + p_v(v-D_S)^2 + \sum_{x\in S\backslash\{v\}}\frac{p_{\sim v}}{|S|-1}(x-D_S)^2}\right] \quad (2)$$

$$= \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}\left[\frac{2(X_t-C_S)(X_{t+\tau}-C_S)}{\frac{1}{|S|}\sum_{x\in S}(x-\mu_u)^2 + p_v(v-D_S)^2 + \sum_{x\in S\backslash\{v\}}\frac{p_{\sim v}}{|S|-1}(x-D_S)^2}\right]. \quad (3)$$

Substituting $S = \{1,2,3,4,5,6\}$ into Equation 2 gives us our result. $\square$

It is then easy to construct examples by substituting values for v and notice that, irrespective of v if $p_v = \frac{1}{6}$ (*i.e.* the biased die is actually unbiased), μ reduces to $\frac{7}{2} = \mu_u$, and σ^2 reduces to $\frac{35}{12} = \sigma_u^2$, as expected.

This analysis has related the learning rates of linear classifiers such as Logistic Regression (LR) of the decay in correlations in the sequence, which motivates the empirical use of auto-correlation as a sensible quantity to estimate when deciding whether or not a structured model is required. There are two main factors affecting the decay of correlations in a sequence: the strength of the chaos in the underlying dynamical system $g : X \to X$, and the *regularity* of the observables F and G. Generally speaking, the correlations decay rapidly if the system is strongly chaotic and the observations are sufficiently regular (*e.g.* systems that are Hölder continuous. We shall see that many real-world problems that have been considered to be sequential classification tasks, and hence 'requiring' structured models, in fact do exhibit the rapid decays in auto-correlation required by the theory.

References

Priestley(1988). Priestley MB (1988) Non-Linear and Non-Stationary Time Series Analysis. Academic Press