Blackwell Prediction for Categorical Data

H. R. Lerche
University of Freiburg i. Br.

Abstract

We study the problem of sequential prediction of categorical data and discuss a generalisation of Blackwell's algorithm on 0-1 data. The arguments are based on Blackwell's approachability results given in [1]. They use mainly linear algebra.

1 Introduction and Background

Let us consider the problem of sequential prediction of categorical data. Let $D = \{0, 1, \dots, d-1\}$ denote the set of possible outcomes with $d \geq 2$. Let $x_1, x_2, \dots$ be an infinite sequence with values in D. Let $Y_1, Y_2, \dots$ denote the sequence of predictions. This is a random sequence with values in D. Y_{n+1} predicts x_{n+1} and may depend on the first n outcomes $x_1, x_2, \dots, x_n, Y_1, Y_2, \dots, Y_n$ and some additional random mechanism. Our goal ist to construct a sequential prediction procedure which works well for all sequences $(x_i)_{i\in\mathbb{N}}$ in an asymptotic sense. We intend to generalize Blackwell's prediction procedure for two categories. The algorithm of Blackwell can be described as follows using Figure 1 below. Let $x_1, x_2, \dots$ be an infinite 0-1 sequence. Let $\overline{x}_n = \frac{1}{n}\sum_{k=1}^n x_k$ be the relative frequency of the "ones" and $\overline{\gamma}_n = \frac{1}{n}\sum_{k=1}^n \mathbb{1}_{\{Y_k=x_k\}}$ the relative frequency of correct guesses. Let $\mu_n = (\overline{x}_n, \overline{\gamma}_n) \in [0,1]^2$ and $\mathcal{S} = \{(x,y) \in [0,1]^2 \mid y \geq \max(x, 1-x)\}$.

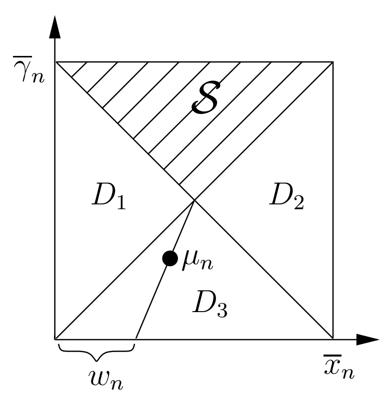

Figure 1

[0] *Date:* 15 Feb 2011

In Fig. 1, let D_1, D_2 and D_3 be the left, right, and bottom triangles, respectively, in the unit square so that $D_1 = \{(x, y) \in [0, 1]^2 \mid x \le y \le 1-x\}$ etc. When $\mu_n \in D_3$, draw the line through the points μ_n and $(\frac{1}{2}, \frac{1}{2})$ and let $(w_n, 0)$ be the point where this line crosses the horizontal axis. The Blackwell algorithm chooses its prediction Y_{n+1} on the basis of μ_n according to the (conditional) probabilities

$$P(Y_{n+1} = 1) = \begin{cases} 0 & \text{if } \mu_n \in D_1 \\ 1 & \text{if } \mu_n \in D_2 \\ w_n & \text{if } \mu_n \in D_3. \end{cases}$$

When μ_n is in the interior of $\mathcal{S}$, Y_{n+1} can be chosen arbitrarily. Let $Y_1 = 0$. It then holds that for the Blackwell algorithm applied to any 0-1 sequence $x_1, x_2, \ldots$ the sequence $(\mu_n;\ n \ge 1)$ converges almost surely to $\mathcal{S}$, i.e. $\text{dist}(\mu_n, \mathcal{S}) \to 0$ as $n \to \infty$ almost surely. Here $\text{dist}(\cdot, \cdot)$ denotes the Euclidean distance from μ_n to $\mathcal{S}$.

As Blackwell once pointed out this is a direct consequence of his Theorem 1 in [1] when one chooses the payoff matrix as

$$\begin{pmatrix} (0,1) & (1,0) \\ (0,0) & (1,1) \end{pmatrix}.$$

For a quick almost sure argument see [4]. Blackwell also raised the question whether his Theorem 1 of [1] applies to sequential prediction when there are more than two categories. We shall study this question and finally answer it affirmative.

We construct a Blackwell type prediction procedure for $d > 2$ categories by choosing the state space and the randomisation rules in a certain way. This procedure then has similar properties as Blackwell's original one. It also has the feature that the d-category procedure reduces to the $(d-1)$ category procedure if one category is not observed.

The structure of this paper is as follows. In Section 2 we introduce the appropriate state space and define the randomisation rule. In Section 3 we state the convergence result and prove it. For that we shall apply a simplified version of Blackwell's Theorem 1 of [1], which we also state in Section 3.

This paper is a continuation of [2], where the case $d = 3$ was discussed, and of the diploma thesis of R. Sandvoss [5].

We shall use the following notation: Latin letters for points, vectors, and indices, greek letters for scalars. We denote components of vectors or points by superindices like $v = (v^{(0)}, \ldots, v^{(d-1)}) \in \mathbb{R}^d$. $e_0 = (1, 0, \ldots, 0), \ldots, e_{d-1} = (0, \ldots, 0, 1)$ denote the d-dimensional unit points and $\mathbb{1}_d = (1, \ldots, 1)$. The affine subspace of $\mathbb{R}^d$ generated by the points $a_0, \ldots, a_n \in \mathbb{R}^d$ is given by

$$A(\{a_0, \ldots, a_n\}) := \left\{ a \in \mathbb{R}^d \,\middle|\, a = \sum_{i=0}^{n} \lambda_i a_i, \sum_{i=0}^{n} \lambda_i = 1, \lambda_i \in \mathbb{R}, a_i \in \mathbb{R}^d, i = 0, \ldots, n \right\}.$$

The convex hull of $a_i, \dots, a_n \in \mathbb{R}^d$ is given by

$$\begin{aligned}&\operatorname{conv}(\{a_0, \dots, a_n\})\\&\quad= \left\{ a \in \mathbb{R}^d \,\middle|\, a = \textstyle\sum_{i=0}^n \lambda_i a_i, \sum_{i=0}^n \lambda_i = 1, \lambda_i \in [0,1], a_i \in \mathbb{R}^d, i = 0, \dots, n \right\}.\end{aligned}$$

The Euclidean scalar product on $\mathbb{R}^d$ is given by $\langle \cdot, \cdot \rangle$, the Euclidean distance by $\operatorname{dist}(\cdot, \cdot)$.

2 The Construction of the d-Dimensional Prediction Procedure

2.1 The Structure of the Prediction Prism

For $n \in \mathbb{N}$, $x_1, x_2, \dots, x_n \in D$ let $Y_1, Y_2, \dots, Y_n \in D$ denote the corresponding predictions. Let $\overline{x}_n = (\overline{x}_n^{(0)}, \dots, \overline{x}_n^{(d-1)})$ with $\overline{x}_n^{(l)} = \frac{1}{n} \sum_{k=1}^n \mathbb{1}_{\{x_i = l\}}$, $l \in D$, denote the vector of the relative frequencies of the n outcomes and $\overline{\gamma}_n = \frac{1}{n} \sum_{k=1}^n \mathbb{1}_{\{Y_k = x_k\}}$ the relative frequency of correct predictions.

Let

$$\mathbf{\Sigma}_{d-1} = \left\{ (q_0, \dots, q_{d-1}) \mid q_l \geq 0, \ \sum_{l=0}^{d-1} q_l = 1 \right\}$$

denote the unit simple in $\mathbb{R}^d$ and

$$W_d = \mathbf{\Sigma}_{d-1} \times [0,1] = \{ (q, \gamma) \mid q \in \Sigma_{d-1}, \ 0 \leq \gamma \leq 1 \}.$$

Since $\sum_{l=0}^{d-1} x_n^{(l)} = 1$, we have $\overline{x}_n \in \mathbf{\Sigma}_{d-1}$ and $(\overline{x}_n, \overline{\gamma}_n) \in W_d$. Let $\mathscr{S}_d = \{ (q, \gamma) \in W_d \mid \gamma \geq \max_l q^{(l)} \}$. We are interested in prediction procedures for which $\mu_n := (\overline{x}_n, \overline{\gamma}_n)$ converges to $\mathscr{S}_d$ for every sequence $x_1, x_2, \dots$ This means that the Euclidean distance $\operatorname{dist}(\mu_n, \mathscr{S}_d) \to 0$ as $n \to \infty$.

Unfortunately Blackwell's Theorem 1 of [1] cannot be applied directly. The reader may take a look at Theorem 3.3 below which is a simplified version of Blackwell's result. The condition (C) there does not hold in general for W_d and $\mathcal{S}_d$. (To see this, let $d = 3$, $s = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, $\mu_n = (\frac{1}{4}, \frac{1}{4}, \frac{1}{2}, 0)$. Then $p(\mu_n) = \mu_n$, and $s - \mu_n$ is not perpendicular to $\mathcal{R}(p(\mu_n))$.)

The difficulties vanish when one modifies the state space in the right way. Let $V_d = \{ q + \gamma \mathbb{1}_d \mid (q, \gamma) \in W_d \}$ with $\mathbb{1}_d = (1, \dots, 1)$. Then $v_n := \overline{x}_n + \overline{\gamma}_n \mathbb{1}_d \in V_d$ for all n. The convergence of μ_n to $\mathscr{S}_d$ corresponds to that of v_n to $\boldsymbol{\mathcal{S}}_d$ where $\boldsymbol{\mathcal{S}}_d = \{ q + \gamma \mathbb{1}_d \in V_d \mid \gamma \geq \max_l q^{(l)} \}$. This follows from the fact that $\Psi : W_d \to V_d$ with $\Psi((q, \gamma)) = q + \gamma \mathbb{1}_d$ is an isometric bijection of W_d on V_d. We note that for $z, z' \in W_d$ it holds that

$$\operatorname{dist}(\Psi(z), \Psi(z'))^2 = \sum_{i=0}^{d-1} (z_i - z_i')^2 + d \cdot (z_d - z_d')^2.$$

To construct the appropriate randomisation regions let us "cut" the prism V_d by certain hyperplanes. (This corresponds to splitting the unit square by the diagonals in the case of two categories.)

Let $e_0 = (1, 0, 0, \ldots, 0), \ldots, e_{d-1} = (0, 0, \ldots, 0, 1)$ denote the d-dimensional unit points. Let $E_l = A(\{e_0, \ldots, e_{l-1}, e_l + \mathbb{1}_d, e_{l+1}, \ldots, e_{d-1}\})$, $l = 0, \ldots, d-1$, denote the hyperplanes which contain one vertex of the "upper side" of the prism $e_l + \mathbb{1}_d$ and $(d-1)$ vertices $e_k \neq e_l$ of $\boldsymbol{\mathcal{S}}_{d-1}$. The d hyperplanes E_l cut the prism V_d in 2^d pieces, and all contain the point $s = (\frac{2}{d}, \frac{2}{d}, \ldots, \frac{2}{d})$. In this point s the planes E_l are all perpendicular to each others.

This can easily be seen since their corresponding normal vectors are given by $n_l = -e_l + \frac{2}{d}\mathbb{1}_d$. This leads to the following characterization of lying "above" E_i:

$$v \text{ lies above } E_i \Leftrightarrow \langle v - n_i, n_i \rangle < 0.$$

In the same way one defines lying below and in E_i.

Now we can describe $\boldsymbol{\mathcal{S}}_d$ in two different ways:

$$\begin{aligned}\mathcal{S}_d &= \{q + \gamma \mathbb{1}_d \in V_d \mid \langle q - n_l, n_l \rangle \geq 0 \text{ for } l = 0, \ldots, d-1\} \\ &= \{q + \gamma \mathbb{1}_d \in V_d \mid \gamma \geq \max(q^{(0)}, \ldots, q^{(d-1)})\}.\end{aligned}$$

For the case $d = 3$ the sets V_d and $\mathcal{S}_d$ are shown in the following figures.

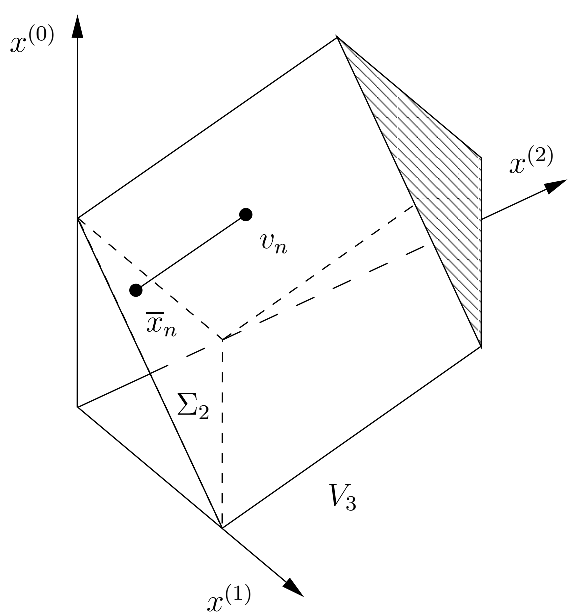

Figure 2

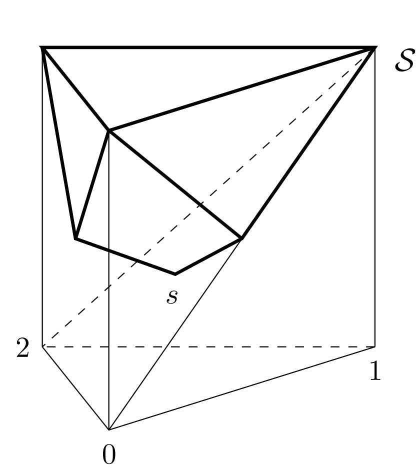

Figure 3

2.2 The Randomisation Rule

For $v_n = \overline{x}_n + \overline{\gamma}_n \mathbb{1}_d$ we will define a d-dimensional random vector $p(v_n) \in \mathbf{\Sigma}_{d-1}$. It plays the same role as w_n does in the 0-1 case. With it we define Y_{n+1} :

$$P(\{Y_{n+1} = k\}) = p^{(k)}(v_n) \text{ for } k \in D.$$

Definition 2.1 *Let $v_n \in V_d$, $n \in \mathbb{N}$ and let $(i_0, \ldots, i_{d-1})$ be a permutation of $(0, \ldots, d-1)$ such that it holds:*

$$\begin{aligned} &\langle v_n - n_l, n_l \rangle \leq 0 \quad \text{for } l = i_0, \ldots, i_j \\ \text{and} \ &\langle v_n - n_l, n_l \rangle > 0 \quad \text{for } l = i_{j+1}, \ldots, i_{d-1}. \end{aligned}$$

Case 1: Let $v_n \in V_d \setminus \mathcal{S}_d$.

Let $A_1 = A(\{\frac{2}{d}\mathbb{1}_d, e_{i_{j+1}}, \ldots, e_{i_{d-1}}, v_n\})$ be the affine space of $\mathbb{R}^d$ generated by the points in the waved brackets. Let $A_2 = A(\{e_{i_0}, \ldots, e_{i_j}\})$ denote the corresponding affine space. The intersection $A_1 \cap A_2$ contains exactly one point of $\mathbf{\Sigma}_{d-1}$, we call it $p(v_n)$.

Case 2: Let $v_n \in \partial\mathcal{S}_d$. Let $\nu = \#\{E_k \mid v_n \in E_k \text{ for } k = 0, \ldots, d-1\}$.

Then

$$p^{(k)}_{(v_n)} = \begin{cases} 1/\nu & \text{for } v_n \in E_k \\ 0 & \text{for } v_n \notin E_k \end{cases}$$

for $k = 0, 1, \ldots, d-1$.

The prediction procedure just defined is called "Generalized Blackwell algorithm".

Remarks 2.2 *1) The case $v_n \in \mathcal{S}_d \setminus \partial\mathcal{S}_d$ does not occur by the construction of the rule.*

2) $A_2 = \emptyset$ cannot occur, since then there exists at least one $k \in D$ with $\langle v_n - n_k, n_k \rangle \leq 0$.

3) We note that $A_1 \cap A_2$ contains always just one point of $\mathbf{\Sigma}_{d-1}$.

4) For $j = d-1$ one obtains $A_1 = A(\{\frac{2}{d}\mathbb{1}_d, v_n\})$, $A_2 = A(\{e_{i_0}, \ldots, e_{i_{d-1}}\})$ and $p(v_n)$ is the projection along the line, defined by $\frac{2}{d}\mathbb{1}_d$ and v_n "down" to Σ_{d-1}.

5) For $d = 3$ the following figure shows the randomisation in a "lower" side piece of the prism. Here planes lie above μ_n and one below.

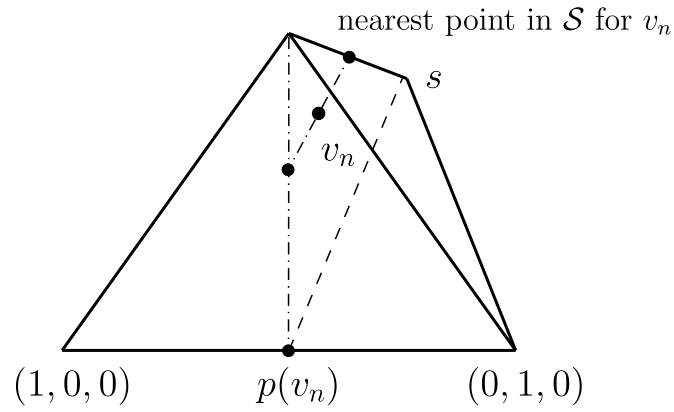

Figure 4

3 The Convergence Result

3.1 Main Result

Theorem 3.1 *Let $d \geq 2$. Then for the generalized Blackwell algorithm, applied to any infinite sequence $x_1, x_2, \ldots$ with values in D, it holds that* $\operatorname{dist}(v_n, \mathcal{S}_d) \to 0$ *with probability one as $n \to \infty$.*

Now we shall derive Theorem 3.1 by tracing it back to Blackwell's Theorem 1 of [1]. This we first state in a simplified version.

3.2 Blackwell's Minimax Theorem

We consider a repeated game of two players with a payoff matrix $M = (m_{ij})$ with $m_{ij} \in \mathbb{R}^d$ and $1 \leq i \leq r$ and $1 \leq j \leq s$. Player I chooses the row, player II the column. Let

$$\mathcal{P} = \left\{ p = (p_1, \ldots, p_r) \,\middle|\, p_i \geq 0, \sum_{i=1}^{r} p_i = 1 \right\}$$

denote the mixed actions of player I and

$$\mathcal{Q} = \left\{ q = (q_1, \ldots, q_s) \,\middle|\, q_j \geq 0, \sum_{j=1}^{s} q_j = 1 \right\}$$

the mixed actions of player II. A strategy f in a repeated game for player I is a sequence $f = (f_k;\ k \geq 1)$ with $f_k \in \mathcal{P}$. A strategy g for player II is defined similarly. Two strategies define a sequence of payoffs z_k, $k = 1, 2, \ldots$ In detail: If in the k-th game i and j are choosen according to f_k and g_k, the payment to player I is $m_{ij} \in \mathbb{R}^d$. Blackwell discussed in [1] the question: Can player I control

$\overline{z}_n = \frac{1}{n}\sum_{k=1}^{n} z_k$ with a certain strategy such that $\overline{z}_n$ approaches a given set $\mathcal{S}$ independently of what player II does?

Definition 3.2 *A set $\mathcal{S} \subset \mathbb{R}^d$ is approachable for player I if there exists a strategy f^* for which $\operatorname{dist}(\overline{z}_n, \mathcal{S}) \to 0$ with probability one.*

Theorem 3.3 (Blackwell) *For $p \in \mathcal{P}$ let*

$$\mathcal{R}(p) = \operatorname{conv}\left(\sum_{i=1}^{r} p_i m_{ij};\ j = 1, 2, \ldots, s\right).$$

Let $\mathcal{S}$ denote a closed convex subset of $\mathbb{R}^d$. For every $z \notin \mathcal{S}$ let y denote the closest point in $\mathcal{S}$ to z. We assume:

(C) *For every $z \notin \mathcal{S}$ there exists a $p(z) \in \mathcal{P}$ such that the hyperplane through y, which is perpendicular to the line segment $\overline{zy}$, seperates z from $\mathcal{R}(p(z))$.*

Then $\mathcal{S}$ is approachable for player I.

3.3 Proof of the Main Result

To apply Theorem 3.3 to our case, we choose the vertices of V_d as "payments":

$$m_{ij} = \begin{cases} e_i + \mathbb{1}_d & \text{if } i = j, \\ e_j & \text{if } i \neq j. \end{cases}$$

We choose $\mathcal{S}$ as $\mathcal{S}_d = \{q + \gamma \mathbb{1}_d \in V_d \mid \gamma \geq \max_l q^{(l)}\}$. Then

$$\begin{aligned}
\mathcal{R}(p) &= \operatorname{conv}\left(\left\{ \sum_{i=0, i\neq j}^{d-1} p^{(i)} e_j + p^{(j)}(e_j + \mathbb{1}_d) \,\middle|\, j = 0, \ldots, d-1 \right\}\right) \\
&= \operatorname{conv}\left(\left\{ \sum_{i=0}^{d-1} p^{(i)} e_j + p^{(j)} \mathbb{1}_d \,\middle|\, j = 0, \ldots, d-1 \right\}\right) \\
&= \operatorname{conv}\big(\{ e_j + p^{(j)} \mathbb{1}_d \mid j = 0, \ldots, d-1 \}\big).
\end{aligned}$$

It is left to show that condition (C) is fulfilled.

Let $v \in V_d \setminus \mathcal{S}_d$. We denote by v_{proj} the closest point in $\mathcal{S}_d$ to v. We will show:

Fact 1 $v_{\text{proj}} \in \mathcal{R}(p(v))$

Fact 2 $v - v_{\text{proj}}$ is perpendicular to $A(\mathcal{R}(p))$. Here $A(\mathcal{R}(p))$ means the smallest affine subspace which contains $\mathcal{R}(p)$.

Both facts together imply condition (C) and finally Theorem 3.1.

For the proofs we shall assume that the following situation holds: For $v \in V_d \setminus \mathcal{S}_d$ it holds

$$\begin{aligned} \langle v - n_i, n_i \rangle &\le 0 \quad \text{for } i = 0, \dots, j \\ \text{and} \quad \langle v - n_i, n_i \rangle &> 0 \quad \text{for } i = j+1, \dots, d-1. \end{aligned}$$

Proof of Fact 1: v lies below E_i for $i = 0, 1, \dots, j$, but $v_{\text{proj}} \in \mathcal{S}_d$. Thus $v_{\text{proj}} \in E_0 \cap \dots \cap E_j$. Then

$$E_0 \cap \dots \cap E_j = A\Big(\Big\{e_{j+1}, \dots, e_{d-1}, \frac{2}{d}\mathbb{1}_d\Big\}\Big).$$

Thus

$$\begin{aligned} v_{\text{proj}} \in & A\Big(\Big\{e_{j+1}, \dots, e_{d-1}\frac{2}{d}\mathbb{1}_d\Big\}\Big) \cap V_d \\ & \subset A\big(\big\{e_i + p^{(i)}(v)\mathbb{1}_d \mid i = 0, \dots, d-1\big\}\big) \cap V_d = \mathcal{R}(p(v)). \end{aligned}$$

The inclusion follows since $p^{(l)}(v) = 0$ for $j+1 = l \le d-1$ and $\frac{2}{d}\mathbb{1}_d = \frac{1}{d}\sum_{i=0}^{d-1}(e_i + p^{(i)}\mathbb{1}_d)$. □

Fact 2 will be proven by a sequence of lemmata. At first we generate a new auxiliary point $\tilde{v}$ which lies in the same plane as $p(v)$.

Lemma 3.4 *For $v \in V_d \setminus \mathcal{S}_d$ let $A' = A(\{v, v_{\text{proj}}\})$ and $A'' = A(\{e_{j+1}, \dots, e_{d-1}, p(v)\})$. Then there exists exactly one point $\tilde{v} \in A' \cap A''$ and $\tilde{v} \notin S_d$.*

Proof: Let $A_1 = A(\{e_{j+1}, \dots, e_{d-1}, \frac{2}{d}\mathbb{1}_d, v\})$ as in Definition 2.1. Then according to Definition 2.1 $p(v) \in A_1$ and $v_{\text{proj}} \in A_1$ by the proof of Fact 1. Then it follows that $\frac{2}{d}\mathbb{1}_d \in A' \vee A''$. Here $A' \vee A''$ denotes the smallest affine space, which contains A', A''. It holds $A_1 = A' \vee A''$. Since A' and A'' are not parallel it follows that $A' \cap A'' \neq \emptyset$ and by the dimension formula $\dim(A' \cap A'') = 0$. Hence $A' \cap A''$ contains exactly one point. We call it $\tilde{v}$. If $\tilde{v} \in \mathcal{S}_d$, then $\tilde{v} \in \mathcal{S}_d \cap A''$. Then $\mathcal{S}_d \cap A(\mathbf{\Sigma}_{d-1}) \neq \emptyset$, which is a contradiction to the definitions of $\mathcal{S}_d$ and $\mathbf{\Sigma}_{d-1}$. □

A direct consequence of Lemma 3.4 is

Fact 3: a) $v_{\text{proj}} = (\tilde{v})_{\text{proj}}$;

b) $v - v_{\text{proj}} \perp A(\mathcal{R}(p)) \Leftrightarrow \tilde{v} - (\tilde{v})_{\text{proj}} \perp A(\mathcal{R}(p))$.

We shall use Fact 3 to show Fact 2. At first we calculate $(\tilde{v})_{\text{proj}}$ from $\tilde{v}$. For simplification, we write $\tilde{v}_{\text{proj}}$ instead of $(\tilde{v})_{\text{proj}}$ from now on.

Lemma 3.5

$$\tilde{v}^{(l)}_{\text{proj}} = \begin{cases} \frac{2}{d}\Bigg(1 - \sum\limits_{k=j+1}^{d-1} \lambda_k\Bigg) & \textit{for } l = 0, \dots, j, \\ \frac{2}{d}\Bigg(1 - \sum\limits_{\substack{k=j+1 \\ k \neq l}}^{d-1} \lambda_k\Bigg) + \big(1 - \frac{2}{d}\big)\lambda_l & \textit{for } l = j+1, \dots, d-1, \end{cases}$$

where $\tilde{v} = p + \lambda_{j+1}(e_{j+1} - p) + \cdots + \lambda_{d-1}(e_{d-1} - p) \in A''$.

Proof: From the proofs of Fact 1 and 3 it follows that

$$\tilde{v}_{\text{proj}} \in A(\{e_{j+1}, \dots, e_{d-1}, \tfrac{2}{d}\mathbb{1}_d\}) \cap \mathcal{S}_d.$$

The smallest affine space, which contains this set is given by

$$A = \left\{a \in \mathbb{R}^d \mid a = \tfrac{2}{d}\mathbb{1}_d + \delta_{j+1}(e_{j+1} - \tfrac{2}{d}\mathbb{1}_d) + \ldots + \delta_{d-1}(e_{d-1} - \tfrac{2}{d}\mathbb{1}_d)\right\}.$$

To find $\tilde{v}_{\text{proj}}$ the projection for v on $\mathcal{S}_d$, we minimize the distance of v to A.

For $a \in A$

$$d(\tilde{v}, a)^2 = \sum_{l=0}^{j} \left(\tilde{v}^{(l)} - \frac{2}{d} + \delta_{j+1}\frac{2}{d} + \ldots + \delta_{d-1}\frac{2}{d}\right)^2 + \sum_{l=j+1}^{d-1} \left(\tilde{v}^{(l)} - \frac{2}{d} - \delta_l\left(1 - \frac{2}{d}\right) + \sum_{\substack{k=j+1 \\ k \neq l}}^{d-1} \delta_k \frac{2}{d}\right)^2. \tag{3.1}$$

Calculating partial derivatives with respect to δ_i, $i = j+1, \dots, d-1$, yields

$$\begin{aligned}\frac{\partial d(\tilde{v},a)^2}{\partial \delta_i} &= \sum_{l=0}^{j} 2\left(\tilde{v}^{(l)} - \frac{2}{d} + \delta_{j+1}\frac{2}{d} + \ldots + \delta_{d-1}\frac{2}{d}\right)\frac{2}{d} \\ &\quad + \sum_{l=j+1}^{d-1} 2\left(\tilde{v}^{(l)} - \frac{2}{d} - \delta_l\left(1-\frac{2}{d}\right) + \sum_{\substack{k=j+1\\ k\neq l}}^{d-1} \delta_k \frac{2}{d}\right)\alpha \\ &= 2\left(\frac{2}{d}\sum_{\substack{l=0\\ l\neq i}}^{d-1} \tilde{v}^{(l)} - \left(1-\frac{2}{d}\right)\tilde{v}^{(i)} - \frac{2}{d} + \delta_i\right),\end{aligned}$$

where $\alpha = \frac{2}{d}$ for $l \neq i$, $\alpha = -(1-\frac{2}{d})$ for $l = i$, and thus

$$\frac{\partial d(\tilde{v},a)^2}{\partial \delta_i} = 0 \;\Leftrightarrow\; \delta_i = \frac{2}{d}\left(1 - \sum_{\substack{l=0\\ l\neq i}}^{d-1} \tilde{v}^{(l)}\right) + \left(1-\frac{2}{d}\right)\tilde{v}^{(i)}.$$

The determinant of the Hessian is positive which shows that a minimum occurs. According to the statement of Lemma 3.5 the components of $\tilde{v}$ has the following representation

$$\tilde{v}^{(l)} = \begin{cases} p^{(l)}\left(1 - \sum_{k=j+1}^{d-1} \lambda_k\right) & \text{for } l = 0,\ldots,j, \\ \lambda_l & \text{for } l = j+1,\ldots,d-1, \end{cases} \tag{3.2}$$

where one should note that $p^{(j+1)} = \ldots = p^{(d-1)} = 0$.

Plugging in the equation of δ_i, $i = j+1,\ldots,d-1$, and noting that $\sum_{l=0}^{j} p^{(l)} = 1$ leads to

$$\delta_i = \frac{2}{d}\left(1 - \sum_{l=0}^{j} p^{(l)}\left(1 - \sum_{k=j+1}^{d-1}\lambda_k\right) - \sum_{\substack{l=j+1\\ l\neq i}}^{d-1}\lambda_k\right) + \left(1-\frac{2}{d}\right)\lambda_i$$

and finally to $\delta_i = \lambda_i$. Plugging this in equation (3.1) leads to the statement of the Lemma. □

Lemma 3.6 *It holds:*

1)

$$(\tilde{v} - \tilde{v}_{\text{proj}})^{(l)} = \begin{cases} \left(p^{(l)} - \frac{2}{d}\right)\left(1 - \sum_{k=j+1}^{d-1}\lambda_k\right) & \textit{for } l = 0,\ldots,j, \\ -\frac{2}{d}\left(1 - \sum_{k=j+1}^{d-1}\lambda_k\right) & \textit{for } l = j+1,\ldots,d-1. \end{cases} \tag{3.3}$$

2) The smallest affine subspace which contains $\mathcal{R}(p)$ can be expressed as $x + U$ where one can choose $x = \tilde{v}_{\text{proj}}$ and

$$\begin{array}{ll} e_i + p^{(i)}\mathbb{1}_d - \tilde{v}_{\text{proj}} & \text{for } i = 0, \ldots, j \\ e_i - \tilde{v}_{\text{proj}} & \text{for } i = j+1, \ldots, d-1 \end{array}$$

as linear generating system of U.

Proof: Statement 1) is a direct consequence of Lemma 3.5 and (3.1). Statement 2) follows from the fact that $\tilde{v}_{\text{proj}} = v_{\text{proj}} \in \mathcal{R}(p(v))$ and that $\mathcal{R}(p) = \text{conv}(e_i + p^{(i)}\mathbb{1}_d \mid i = 1, \ldots, d-1)$ where $p^{(j+1)} = \ldots = p^{(d-1)} = 0$.

Lemma 3.7 *It holds*

$$\tilde{v} - \tilde{v}_{\text{proj}} \perp e_i + p^{(i)}\mathbb{1}_d - \tilde{v}_{\text{proj}} \quad \text{for } i = 0, \ldots, j.$$

Proof: Lemma 3.5 implies

$$\begin{aligned} &\left(e_i + p^{(i)}\mathbb{1}_d - \tilde{v}_{\text{proj}}\right)^{(l)} \\ &= \begin{cases} p^{(i)} - \frac{2}{d}\left(1 - \sum_{k=j+1}^{d-1} \lambda_k\right) & l = 0, \ldots, j;\ l \neq i, \\ 1 + p^{(i)} - \frac{2}{d}\left(1 - \sum_{k=j+1}^{d-1} \lambda_k\right) & l = i, \\ p^{(i)} - \frac{2}{d}\left(1 - \sum_{k=j+1, k\neq l}^{d-1} \lambda_k\right) - \left(1 - \frac{2}{d}\right)\lambda_l & l = j+1, \ldots, d-1. \end{cases} \end{aligned} \tag{3.4}$$

From (3.3) and (3.4) it follows

$$
\begin{aligned}
&\langle \tilde v - \tilde v_{\text{proj}}, e_i + p^{(i)}\mathbb{1}_d - \tilde v_{\text{proj}}\rangle \\
&\quad = \sum_{\substack{l=0\\ l\neq i}}^{j} \left(p^{(l)} - \frac{2}{d}\right)\left(1 - \sum_{k=j+1}^{d-1}\lambda_k\right)\left(p^{(i)} - \frac{2}{d}\left(1 - \sum_{k=j+1}^{d-1}\lambda_k\right)\right) \\
&\qquad + \left(p^{(i)} - \frac{2}{d}\right)\left(1 - \sum_{k=j+1}^{d-1}\lambda_k\right)\left(1 + p^{(i)} - \frac{2}{d}\left(1 - \sum_{k=j+1}^{d-1}\lambda_k\right)\right) \\
&\qquad - \sum_{l=j+1}^{d-1}\frac{2}{d}\left(1 - \sum_{k=j+1}^{d-1}\lambda_k\right)\left(p^{(i)} - \frac{2}{d}\left(1 - \sum_{k=j+1 k\neq l}^{d-1}\lambda_k\right) - \left(1 - \frac{2}{d}\right)\lambda_l\right) \\
&\quad = \left(1 - \sum_{k=j+1}^{d-1}\lambda_k\right)\cdot\Bigg[\left(p^{(i)} - \frac{2}{d}\right) + \sum_{i=0}^{j} p^{(l)}\left(p^{(i)} - \frac{2}{d}\left(1 - \sum_{k=j+1}^{d-1}\lambda_k\right)\right) \\
&\qquad - \sum_{l=0}^{i}\frac{2}{d}\left(p^{(i)} - \frac{2}{d}\left(1 - \sum_{k=j+1}^{d-1}\lambda_k\right)\right) - \sum_{l=j+1}^{d-1}\frac{2}{d}\left(p^{(i)} - \frac{2}{d}\left(1 - \sum_{k=j+1}^{d-1}\lambda_k\right)\right) \\
&\qquad + \sum_{k=j+1}^{d-1}\frac{2}{d}\frac{2}{d}\lambda_k + \sum_{l=j+1}^{d-1}\frac{2}{d}\left(1 - \frac{2}{d}\right)\lambda_l\Bigg] \\
&\quad = \left(1 - \sum_{k=j+1}^{d-1}\lambda_k\right) \\
&\qquad \cdot\Bigg[p^{(i)} - \frac{2}{d} + p^{(i)} - \frac{2}{d}\left(1 - \sum_{k=j+1}^{d-1}\lambda_k\right) - d\frac{2}{d}\left(p^{(i)} - \frac{2}{d}\left(1 - \sum_{k=j+1}^{d-1}\lambda_k\right)\right) \\
&\qquad\quad + \sum_{k=j+1}^{d-1}\frac{2}{d}\frac{2}{d}\lambda_k + \sum_{l=j+1}^{d-1}\frac{2}{d}\lambda_l - \sum_{l=j+1}^{d-1}\frac{2}{d}\frac{2}{d}\lambda_l\Bigg] = 0. \qquad \square
\end{aligned}
$$

Lemma 3.8 *It holds*

$$\tilde v - \tilde v_{\text{proj}} \quad \perp \quad e_i + \tilde v_{\text{proj}} \qquad \textit{for } i = j+1, \ldots, d-1.$$

Proof: By Lemma 3.5 one gets

$$(e_i - \tilde{v}_{\text{proj}})^{(l)} = \begin{cases} -\frac{2}{d}\left(1 - \sum\limits_{k=j+1}^{d-1} \lambda_k\right) & l = 0, \ldots, j, \\ -\frac{2}{d}\left(1 - \sum\limits_{\substack{k=j+1 \\ k \neq l}}^{d-1} \lambda_k\right) - \left(1 - \frac{2}{d}\right)\lambda_l & l = j+1, \ldots, d-1;\ l \neq i, \\ 1 - \frac{2}{d}\left(1 - \sum\limits_{\substack{k=j+1 \\ k \neq i}}^{d-1} \lambda_k\right) - \left(1 - \frac{2}{d}\right)\lambda_i & l = i. \end{cases} \tag{3.5}$$

From (3.3) and (3.5) it follows

$$\begin{aligned} &\langle \tilde{v} - \tilde{v}_{\text{proj}}, e_i - \tilde{v}_{\text{proj}} \rangle \\ &= \sum_{l=0}^{j} \left(p^{(l)} - \frac{2}{d}\right) \left(1 - \sum_{k=j+1}^{d-1} \lambda_k\right) \left(-\frac{2}{d}\left(1 - \sum_{k=j+1}^{d-1} \lambda_k\right)\right) \\ &\quad + \sum_{\substack{l=j+1 \\ l \neq i}}^{d-1} -\frac{2}{d} \left(1 - \sum_{k=j+1}^{d-1} \lambda_k\right) \left(-\frac{2}{d}\left(1 - \sum_{\substack{k=j+1 \\ k \neq l}}^{d-1} \lambda_k\right) - \left(1 - \frac{2}{d}\right)\lambda_l\right) \\ &\quad - \frac{2}{d}\left(1 - \sum_{k=j+1}^{d-1} \lambda_k\right) \left(1 - \frac{2}{d}\left(1 - \sum_{\substack{k=j+1 \\ k \neq i}}^{d-1} \lambda_k\right) - \left(1 - \frac{2}{d}\right)\lambda_i\right) \\ &= \left(1 - \sum_{k=j+1}^{d-1} \lambda_k\right) \\ &\quad \cdot \Bigg[\sum_{l=0}^{i} p^{(l)} \left(-\frac{2}{d}\left(1 - \sum_{k=j+1}^{d-1} \lambda_k\right)\right) + \sum_{l=0}^{j} \frac{2}{d}\frac{2}{d}\left(1 - \sum_{k=j+1}^{d-1} \lambda_k\right) \\ &\qquad + \sum_{l=j+1}^{d-1} \frac{2}{d}\frac{2}{d}\left(1 - \sum_{k=j+1}^{d-1} \lambda_k\right) + \sum_{k=j+1}^{d-1} \frac{2}{d}\frac{2}{d}\lambda_k + \sum_{l=j+1}^{d-1} \frac{2}{d}\left(1 - \frac{2}{d}\right)\lambda_l - \frac{2}{d}\Bigg] \\ &= \left(1 - \sum_{k=j+1}^{d-1} \lambda_k\right) \\ &\quad \cdot \Bigg[-\frac{2}{d}\left(1 - \sum_{k=j+1}^{d-1} \lambda_k\right) + d\frac{2}{d}\frac{2}{d}\left(1 - \sum_{k=j+1}^{d-1} \lambda_k\right) + \sum_{k=j+1}^{d-1} \frac{2}{d}\frac{2}{d}\lambda_k \\ &\qquad + \sum_{l=j+1}^{d-1} \frac{2}{d}\lambda_l - \sum_{l=j+1}^{d-1} \frac{2}{d}\frac{2}{d}\lambda_l - \frac{2}{d}\Bigg] = 0. \end{aligned}$$

□

Finally we can state the proof of Fact 2: By Lemma 3.6, 3.7, and 3.8 one has $\tilde{v} - \tilde{v}_{\text{proj}} \perp A(\mathcal{R}(p))$. By Fact 1 it follows that $v - v_{\text{proj}} \perp A(\mathcal{R}(p))$. $\square$

Acknowledgements.

References

[1] Blackwell, D. (1956) An Analog of the Minimax Theorem for Vector Payoffs. *Pacific Journal of Mathematics*, **6**, 1–8.

[2] Lerche, H. R., Sakar, J. (1994) The Blackwell Prediction Algorithm for 0-1 sequences and generalization. In *Statistical Decision Theory and Related Topics V*, Eds.: S. S. Gupta, Y. O. Berger, Springer Verlag, 503–511.

[3] Riedel, F. (2008) Blackwell's Theorem with Weighted Averages. Preprint.

[4] Robbins, H. and Siegmund, D. (1971) A Convergence Theorem for Nonnegative Almost Supermartingales and Some Applications. *Optimizing Methods in Statistics*. 233–257. Academic Press, New York.

[5] Sandvoss, R. (1994) Blackwell Vorhersageverfahren – zur Komplexität von Finanzdaten. Diplomarbeit Universität Freiburg.