Probability Vector Estimation under Constraints by Discounting

Günther Wirsching
Hans Fischer
Mathematisch-Geographische Fakultät
Katholische Universität Eichstätt-Ingolstadt

Preprintreihe Mathematik 2011 – 04

Abstract

The focus of this paper is on using observations to estimate an unknown probability vector $p = (p_1, \ldots, p_N)$ supposed to underlie a multinomial process. In some technical applications, e.g., parameter estimation for a hidden Markov chain, numerical stability can be guaranteed only if we assume each estimate $\hat{p}_i$ for a probability p_i conforming to the constraint $\hat{p}_i \geq m$, where $m > 0$ is an appropriate constant depending on the particular technical application. Aiming at such estimates $\hat{p}_i$ we present a fast discounting algorithm which comprises ad-hoc methods known as absolute discounting, linear discounting, and square-root discounting as special cases. In order to base discounting on probabilistic principles, we adopt a Bayesian approach, and we show that, presupposing an arbitrary nonvanishing prior, minimizing the ℓ^∞-norm of a certain risk vector defined by a one-sided loss function leads to a new consistent estimator. It turns out to be quite natural to derive from this an (in general inconsistent) estimator meeting the constraints $\hat{p}_i \geq m$. Using asymptotic statistics, we show that a good approximation to this estimator can be reached by means of our fast discounting algorithm in context with an appropriate adjustment of square-root discounting.

1 Introduction

In this paper we assume that, in accordance with a multinomial process, a fixed number N of mutually exclusive events $E_1, \ldots, E_N$ are produced by unknown probabilities $p_1, \ldots, p_N$, respectively, and that we have performed an experiment with c_0 observations in which each E_i has occurred with a count $c_i \geq 0$ such that $c_0 = \sum_{i=1}^{N} c_i$. Our aim is to use the counts c_i for estimates $\hat{p}_1, \ldots, \hat{p}_N$ of the unknown probabilities, where these estimates are subject to the constraint

$$\forall\, i \in \{1, \ldots, N\} : \quad \hat{p}_i \geq m, \tag{1}$$

m being an appropriate threshold number fixed a priori.

Establishing such a lower bound is motivated by numerical problems arising when Markov chain parameters are estimated from sparse data by use of the so called expectation-maximization (EM) algorithm; typical examples occur in the context of channel modeling in information theory [5, 8], or language modeling [4, 1].

The EM algorithm starts with choosing an appropriate initial Markov matrix M_0. Then it uses a given set of observations, and computes a probability for each observation under the assumption that M_0 describes the process generating the observation. Next the data collected from the observations are used to update the Markov matrix to M_1. After a few steps, this usually leads to a reasonable estimate M_s for the 'true' underlying Markov matrix, provided that initialization and observation data are not too disadvantageous.

For example, let us consider the problem of learning string edit distances [5, 8]. More specifically, suppose that we have observed a situation where a given channel has transformed a given input string $a_1 \ldots a_k$ into an output string $b_1 \ldots b_\ell$, where we assume $k, \ell \lessapprox 30$. The Markov chain model assumes that this transformation results from a composition of *elementary editing operations* like *substitution* of an input character a_i by an output character b_j, *deletion* of an input character, or *insertion* of an output character, where the—possibly context-dependent—probabilities of the different elementary editing operations are the entries of a (sufficiently large) Markov matrix M.

There may be more than one possible sequence of elementary editing operations leading from $a_1 \dots a_k$ to $b_1 \dots b_\ell$, but the Markov matrix M allows to assign to each such sequence an *editing-path probability*, which is just the product of the probabilities of the elementary editing operations used in the editing path. If only substitutions, insertions, and deletions are taken into account, the length of such an editing path is bounded by $n = k+\ell \lessapprox 60$. The total probability that the channel modeled by M transforms $a_1 \dots a_k$ into $b_1 \dots b_\ell$ can than be computed as the sum of editing-path probabilities, extended over all possible editing paths. It is this total probability which—beside a multitude of further data—is needed by the EM algorithm.

For the purposes of numerical stability of the EM algorithm, it appears to be important to avoid zero total probabilities. This is guaranteed when the n-th power of the smallest possible $\hat{p}_i$ ist not smaller than the smallest positive number representable in the software we use on our computer system. For instance, the smallest positive number in double precision is around 10^{-308}; if we have $n \approx 60$, we come to the lower bound

$$\hat{p}_i \geq \sqrt[60]{10^{-308}} \approx 7.36 \cdot 10^{-6},$$

which, in some applications, may be above the smallest relative frequencies that occur. In this paper, we demonstrate our numerical results using a threshold $m = 10^{-5}$.

The most obvious method to obtain estimates $\hat{p}_i$ from counts c_i is to take relative frequencies:

$$\forall\, i \in \{1, \dots, N\} : \quad \hat{p}_i := \frac{c_i}{c_0}.$$

But, if there are counts $c_i = 0$, or, more generally, small counts $c_i < mc_0$, this would conflict with condition (1). Methods for mastering this situation are called *smoothing* or *discounting* methods: Elevation of smaller relative frequencies to m requires that larger relative frequencies have to be *discounted* in order to ensure the stochastic requirement $\sum \hat{p}_i = 1$. A variety of different discounting methods is widely used in the above described context of Markov chain estimation.

The present paper starts with explaining a fast general discounting algorithm which comprises different discounting methods as special cases. These special cases include

(i) *absolute* discounting where the same amount is subtracted from the large relative frequencies [4, p. 216],

(ii) *linear* discounting where an amount proportional to c_i is subtracted from the large relative frequencies [4, p. 216],

(iii) *square-root* discounting where an amount proportional to $\sqrt{c_i(c_0 - c_i)}$ is subtracted from the large relative frequencies [7],

(iv) *modified linear* discounting where an amount proportional to $c_i(c_0 - c_i)$ is subtracted from the large relative frequencies.

For an engineer, there is good reason for square-root discounting: Assume that the occurrence of a certain event E_i has the probability p_i. Then the distribution of relative frequencies c_i/c_0 has (as a consequence of the binomial distribution for c_i) mean $\mu_i := p_i$ and standard deviation $\sigma_i := \sqrt{p_i(1-p_i)/c_0}$. In an engineering context, the standard deviation is often interpreted as the *imprecision* of a measurement of p_i. Hence, the imprecision of a relative frequency c_i/c_0 is

$$\sigma_i = \sqrt{\frac{p_i(1-p_i)}{c_0}} \approx \sqrt{\frac{1}{c_0}\frac{c_i}{c_0}\left(1 - \frac{c_i}{c_0}\right)} = \sqrt{\frac{c_i(c_0 - c_i)}{c_0^3}}.$$

Therefore, square-root discounting means making discounts the larger the more "imprecise" the single relative frequencies are.

In the hitherto discussed discounting methods the respective modifications of relative frequencies c_i/c_0 are each proportional to a function depending only on the count c_i. From a more general point of view, however, it could be possible that the respective "discounts" are functions of all counts $c_1, \dots, c_N$ together. This happens exactly when we are looking for a probabilistic principle which should enable us to determine a more general discounting method that is "best" in a certain sense. In order to achieve this aim we assume a Bayesian prior on the simplex $\Delta^N \subset \mathbb{R}^N$ (which is

the subspace of all probability vectors $(p_1, \ldots, p_N)$), having a strictly positive and continuous density. We prove that minimizing the ℓ^∞-norm of a risk vector obtained by integrating the product of a one-sided vector valued loss function with the Bayesian posterior density over the simplex leads to consistent estimators for the unknown probabilities governing the observations. It is interesting that this consistent estimator already has the property

$$\forall\, i \in \{1, \ldots, N\} : \quad \hat{p}_i > 0.$$

We then observe that minimizing the ℓ^∞-norm of our risk vector amounts to equalizing the different risks encoded in the different components of our risk vector. This enables us to establish a general method of discounting with a prescribed threshold m: In order to gain an optimal estimate meeting requirement (1), we just have to equalize risks as precisely as possible. Note that we do *not* assume that the "true" probabilities have the property $p_i \geq m$. In fact, in the applications mentioned above, we could not subsume this assumption. The requirement (1) for the *estimated values* $\hat{p}_i$ just arises from the necessity of processing the $\hat{p}_i$ in a numerically stable way. Of course, we have to accept the consequence that such an estimator $(\hat{p}_1, \ldots, \hat{p}_N)$ is no longer consistent in situations where some $p_i < m$.

Finally, we provide a connection between our Bayesian investigations and our general discounting algorithm. We not only show that the above mentioned "equalizing risks" method is equivalent to square-root discounting in an asymptotic sense, we even propose an adjustment of square-root discounting to configure our fast discounting algorithm in such a way that it quickly determines a good approximation to the estimates gained by equalizing risks.

2 A Fast General Discounting Algorithm

After having observed counts $c_1, \ldots, c_N$ which add up to c_0, the maximum-linkelihood estimator for the underlying probability vector corresponds to the relative frequencies:

$$\forall\, i \in \{1, \ldots, N\} : \quad \hat{p}_i := \frac{c_i}{c_0}.$$

In order to get the estimates obeying the constraint $\hat{p}_i \geq m$, we have to increase the estimates for indices where the quotient falls below m, and, consequently, decrease the estimates at least with regard to some of the indices with $c_i > mc_0$. Hence, we consider the sets of indices

$$I_0 := \left\{ i \in \{1, \ldots, N\} : \frac{c_i}{c_0} \leq m \right\},$$

and its complement $I_1 := \{1, \ldots, N\} \setminus I_0$. Then we put

$$\forall i \in I_0 : \quad \hat{p}_i := m. \tag{2}$$

The stochastic condition $\sum \hat{p}_i = 1$ can be ensured by absolute discounting, for example. In this case we calculate

$$\alpha := \frac{1}{|I_1|} \left(m \cdot |I_0| + \sum_{i \in I_1} \frac{c_i}{c_0} - 1 \right),$$

and put

$$\forall i \in I_1 : \quad \hat{p}_i := \frac{c_i}{c_0} - \alpha. \tag{3}$$

By (2) and (3), we clearly have

$$\sum_{i=1}^{N} \hat{p}_i = \sum_{i \in I_0} \hat{p}_i + \sum_{i \in I_1} \hat{p}_i = m \cdot |I_0| + \sum_{i \in I_1} \frac{c_i}{c_0} - \alpha \cdot |I_1| = 1,$$

but we can not be sure whether the constraint $\hat{p}_i \geq m$ is fulfilled for all $i \in I_1$. It may be an index $i \in I_1$ with $m < \frac{c_i}{c_0} < m + \alpha$, resulting in $\hat{p}_i < m$. Hence, we would have to iterate the procedure, yielding in a worst case complexity $O(N^2)$. Analogously, the problem of necessary re-iterations may also occur when using linear or square-root discounting.

The following *fast discounting algorithm* starts with ordering the data appropriately and altogether reduces worst case complexity to $O(N \log N)$. The algorithm uses as input the *threshold* m, *initial estimates* $\mu_1, \ldots, \mu_N$, and *discounting bases* $\sigma_1, \ldots, \sigma_N$, and computes a *discounting factor* α such that the estimates are given by

$$\forall i \in \{1, \ldots, N\}: \quad \hat{p}_i := \max\{\mu_i - \alpha\sigma_i, m\}. \tag{4}$$

It runs as follows:

Initialization.

Read parametrization data. Read m, $\mu_1, \ldots, \mu_N$, $\sigma_1, \ldots, \sigma_N$.

Compute. For $i = 1, \ldots, N$ compute $\alpha_i := \dfrac{\mu_i - m}{\sigma_i}$.

Sort indices i such that $\alpha_1 \leq \ldots \leq \alpha_N$.

Compute $M := \sum_{i=1}^{N} \mu_i$ and $S := \sum_{i=1}^{N} \sigma_i$.

Set $\mu_0 := \sigma_0 := 0$, $L := 1 + m$, and $J := 0$.

Repeat

Replace $M \mapsto M - \mu_J$, $\quad S \mapsto S - \sigma_J$, $\quad L \mapsto L - m$.

Set $\alpha := \frac{M-L}{S}$.

Replace J by $J + 1$.

Until $\alpha \leq \alpha_J$.

Estimate. For $i = 1, \ldots, N$ compute $\quad \hat{p}_i := \max\{\mu_i - \alpha\sigma_i, m\}$.

Stop.

We see that the "Estimate"-part of the algorithm is reached after at most N iterations as follows: Assume that the "Until"-condition is not met for $J = 1, \ldots, N-1$. Then updating M, S, and L, leads to $M = \mu_N$, $S = \sigma_N$, and $L = 1 - (N-1)m$. Consequently,

$$\alpha = \frac{M - L}{S} = \frac{\mu_N - m - 1 + Nm}{\sigma_N} < \frac{\mu_N - m}{\sigma_N} = \alpha_N,$$

where the inequality follows from our assumption $Nm < 1$. Then $J = N - 1$ is increased to $J = N$, and the "Until"-condition is satisfied, proving a worst case complexity $O(N \log N)$. If we pool indices i, j when $c_i = c_j$, then worst case time complexity of this algorithm reduces to $O(D \log D)$ where D denotes the number of different counts.

The estimates $\hat{p}_i$ computed by fast discounting are unchanged if the discounting bases σ_i are replaced by $\lambda\sigma_i$ where λ is a fixed positive factor. Indeed, if all σ_i are changed to $\lambda\sigma_i$, then the computed discounting factor changes to α/λ leading to the same $\hat{p}_i$ by formula (4).

We further note that our fast discounting algorithm implements a *weighted least squares method* with constraints and weights $1/\sigma_i$. Formally, it computes the $\hat{p}_i$ such that the expression

$$\sum_{i=1}^{N} \frac{1}{\sigma_i} (\mu_i - \hat{p}_i)^2$$

is minimal, subject to the constraints

$$\sum_{i=1}^{N} \hat{p}_i = 1 \text{ and } \forall 1 \leq i \leq N: \hat{p}_i \geq m.$$

By an appropriate choice of discounting bases σ_i, this general fast discounting algorithm can be configured to perform either absolute or linear or square-root discounting. In each case, choose $\mu_i := \frac{c_i}{c_0}$ for all $i \in \{1, \ldots, N\}$. The choices of the σ_i are as follows:

Absolute discounting: $\sigma_i := 1$.

Linear discounting: $\sigma_i := \mu_i$.

Square-root discounting: $\sigma_i := \sqrt{\mu_i(1-\mu_i)}$,

Modified linear discounting: $\sigma_i := \mu_i(1-\mu_i)$.

By the way, in all these particular cases we have (for $0 \le m < 1/N$)

$$\alpha_i < \alpha_j \Leftrightarrow c_i < c_j.$$

By this fact, arranging the α_i in ascending order is somehow simplified.

3 A Practical Example

We consider the count vector

$$\begin{aligned} c = (&0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,\\ &0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,2,2,2,\\ &4,4,7,18,61,102,104,247,268,278,293,350,426,463,\\ &571,571,572,614,779,815,872,928,965,1095,1224,1288,\\ &1353,1425,1913,1984,2065,2068,2156,2169,2199,2327,\\ &2386,2699,2861,2885,3017,3090,3207,3267,3270,3531,\\ &3804,4300,5413,5781,6413,6504,6534,6768,7240,7481,\\ &7821,7828,8304,8559,8898,9792,10829,11069,12227,\\ &12254,12927,13255,13485,15226,15366,15510,15529,\\ &18587,19937,22791,32288,47562,56644,65832,77061), \end{aligned} \tag{5}$$

with length $N = 131$ and with $c_0 = 662623$. We apply our discounting algorithm for $m = 10^{-5}$, and for absolute, square-root, linear, and modified linear discounting, respectively.

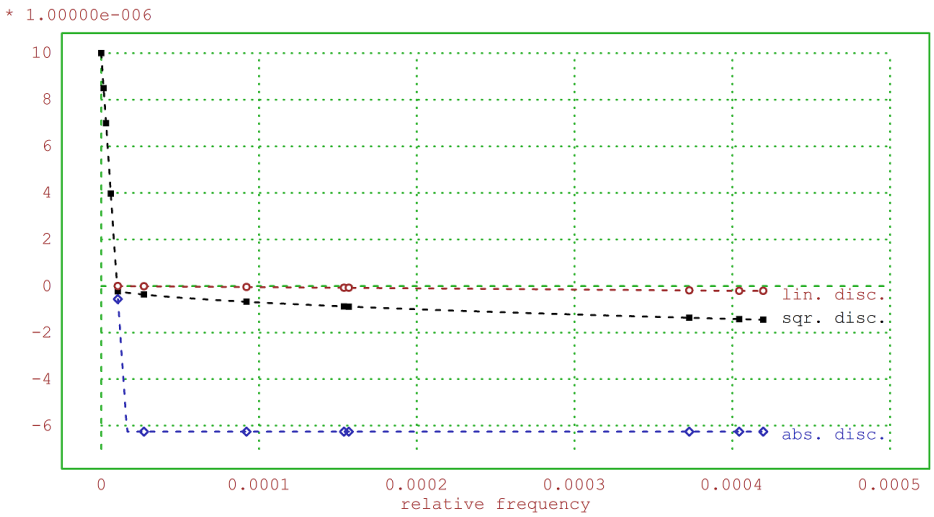

Figure 1: Differences between discounted values and related relative frequencies for small counts as depending on the relative frequencies; the four first values are common to all methods.

In the case of absolute discounting, we have $\mu_i = \frac{c_i}{c_0}$ and $\sigma_i = 1$; in the case of square-root discounting we choose $\mu_i = \frac{c_i}{c_0}$ and $\sigma_i = \max(\sqrt{\mu_i(1-\mu_i)}, \varepsilon)$; in the case of linear discounting we put $\mu_i = \frac{c_i}{c_0}$ and $\sigma_i = \max(\mu_i, \varepsilon)$; finally, in the case of modified linear discounting we put $\mu_i = \frac{c_i}{c_0}$ and $\sigma_i = \max(\mu_i(1-\mu_i), \varepsilon)$, where we choose $\varepsilon := \frac{1}{c_0}$. The particular definition of σ_i in the two latter cases is due to the fact that we have to avoid $\sigma_i = 0$ in the application of the algorithm. In principle we could substitute the lower bound ε of σ_i by any sufficiently small number.

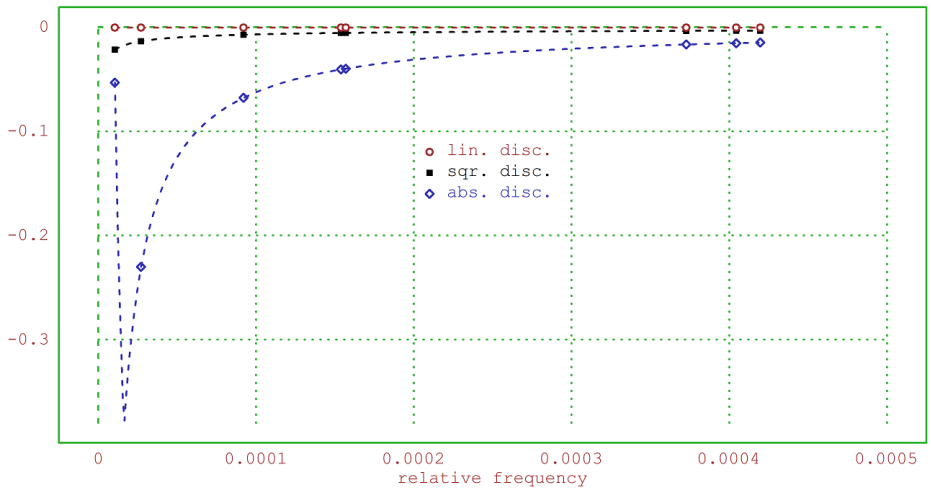

Figure 2: Relative deviations between discounted values and related relative frequencies for small counts as depending on the relative frequencies; differences between linear discounting and modified linear discounting are below the graphical precision.

For small counts the results (up to a precision of $\pm 10^{-8}$) in applying the respective discounting methods are shown in figures 1 and 2 (differences between linear and modified linear discounting are below the precision of these graphics, therefore modified linear discounting is omitted in these figures).

c_i	c_i/c_0	absolute $\hat{p}_i$	square-root $\hat{p}_i$	linear $\hat{p}_i$	mod.linear $\hat{p}_i$
0	0	$1 \cdot 10^{-5}$	$1 \cdot 10^{-5}$	$1 \cdot 10^{-5}$	$1 \cdot 10^{-5}$
1	$1.50 \cdot 10^{-6}$	$1 \cdot 10^{-5}$	$1 \cdot 10^{-5}$	$1 \cdot 10^{-5}$	$1 \cdot 10^{-5}$
2	$3.01 \cdot 10^{-6}$	$1 \cdot 10^{-5}$	$1 \cdot 10^{-5}$	$1 \cdot 10^{-5}$	$1 \cdot 10^{-5}$
4	$6.03 \cdot 10^{-6}$	$1 \cdot 10^{-5}$	$1 \cdot 10^{-5}$	$1 \cdot 10^{-5}$	$1 \cdot 10^{-5}$
7	$1.056 \cdot 10^{-5}$	$1 \cdot 10^{-5}$	$1.033 \cdot 10^{-5}$	$1.055 \cdot 10^{-5}$	$1.055 \cdot 10^{-5}$
18	$2.716 \cdot 10^{-5}$	$2.091 \cdot 10^{-5}$	$2.679 \cdot 10^{-5}$	$2.715 \cdot 10^{-5}$	$2.715 \cdot 10^{-5}$
61	$9.205 \cdot 10^{-5}$	$8.580 \cdot 10^{-5}$	$9.137 \cdot 10^{-5}$	$9.201 \cdot 10^{-5}$	$9.201 \cdot 10^{-5}$
102	0.00015393	0.00014768	0.00015305	0.00015385	0.00015385
⋮	⋮	⋮	⋮	⋮	⋮
278	0.00041954	0.00041329	0.00041809	0.0004193**3**	0.00041932

The global behavior of the differences between discounted values and related relative frequencies can be seen in figure 3.

For $q_i > m$, the relative deviations of discounted values $\hat{p}_i$ and relative frequencies q_i, calculated by the fraction $\frac{\hat{p}_i - q_i}{q_i}$, are depicted in figure 4.

The dashed lines in the figures are inserted to help seeing results from a specific discounting method as connected. They are computed using the fact that, for fixed α, there is an easily computable functional dependency of the quantities in question from relative counts c_i/c_0. (Note that it is not recommended to use such a line for a different count vector $\tilde{c} \neq c$, as α depends on the count vector.)

In comparing the three methods we see that, both for smaller and larger counts, and with regard to absolute deviation $\hat{p}_i - q_i$ as well as with regard to relative deviation $\frac{\hat{p}_i - q_i}{q_i}$, square-root discounting is "between" the results obtained by absolute and linear discounting. Already from this perspective, square-root-discounting seems to be a good "compromise" method. As we will see in the following, square-root discounting can be substantiated by a probabilistic principle, which also yields an even better adaption of this method to estimating probabilities of mutually exclusive events.

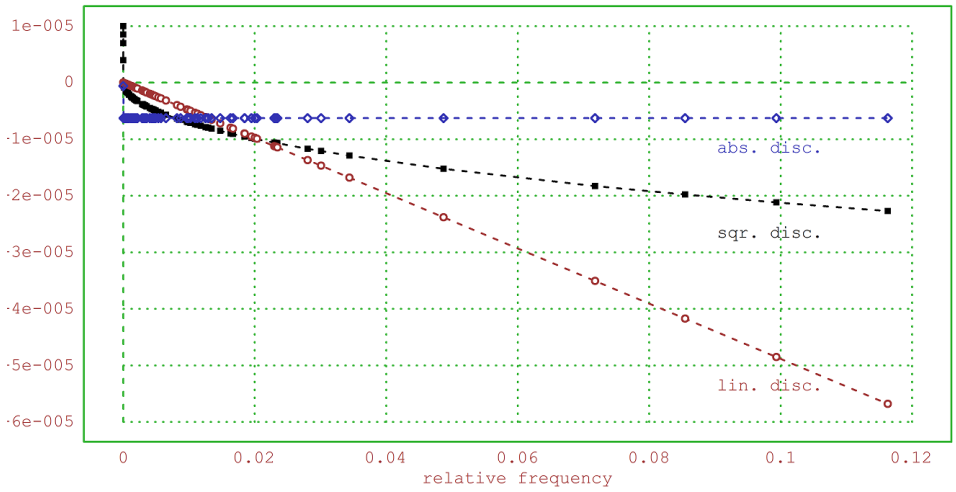

Figure 3: Differences between discounted values and related relative frequencies as depending on the relative frequencies.

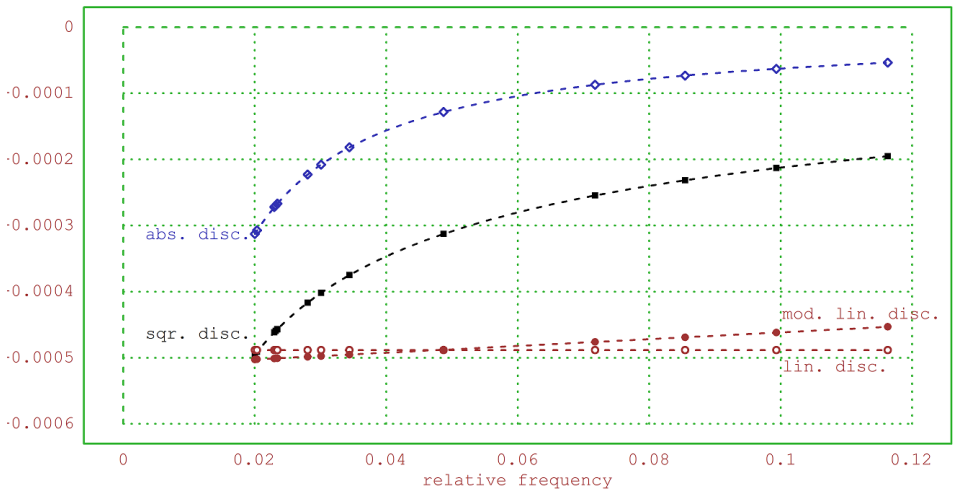

Figure 4: Relative deviations between discounted values and related relative frequencies for larger counts as depending on the relative frequencies.

4 A Bayesian Approach with One-Sided Loss

In the following we use the standard simplex Δ^N in $\mathbb{R}^N$ as the parameter set. This set is defined by

$$\Delta^N := \left\{ (\theta_1, \ldots, \theta_N) \in \mathbb{R}^N \,\middle|\, 0 \le \theta_i \le 1, \sum_{i=1}^{N} \theta_i = 1 \right\}.$$

Given the observations with count vector $c = (c_1, \ldots, c_N)$ and $c_0 := \sum_{i=1}^{N} c_i$, the likelihood function $L : \Delta^N \to \mathbb{R}$ according to a multinomial process is

$$L(\theta) = \frac{c_0!}{c_1! \cdots c_N!} \prod_{i=1}^{N} \theta_i^{c_i}.$$

If the prior is given through a strictly positive, continuous density $\psi : \Delta^N \to \mathbb{R}$, the density of the posterior Π_c is

$$f_c(\theta) = K_c \psi(\theta) \prod_{i=1}^{N} \theta_i^{c_i}, \tag{6}$$

where K_c is a normalization constant determined by the condition that f should be the density of a probability. If $dS(\theta)$ denotes the surface measure on Δ^N, we have

$$K_c = \left(\int_{\Delta^N} \psi(\theta) \prod_{i=1}^{N} \theta_i^{c_i} \, dS(\theta) \right)^{-1}.$$

The *unit-step* or *Heaviside* function is given by

$$u : \mathbb{R} \to \mathbb{R}, \quad \begin{cases} u(t) = 0 & \text{for } t \le 0, \\ u(t) = 1 & \text{for } t > 0. \end{cases} \tag{7}$$

We use it for constructing the *Heaviside loss vector* on the simplex as follows:

$$\ell : \Delta^N \times \Delta^N \to \mathbb{R}^N, \quad \ell_i(x, \theta) = u(\theta_i - x_i) \quad \text{for each } i \in \{1, \ldots, N\}. \tag{8}$$

Integrating the product of Heaviside loss vector and posterior density gives a *risk vector* $r(x)$ with components

$$r_i(x_i) = \int_{\Delta^N} \ell_i(x, \theta) f_c(\theta) \, dS(\theta) = \int_0^1 u(\theta_i - x_i) f_i(\theta_i) \, d\theta_i = \int_{x_i}^1 f_i(\theta_i) \, d\theta_i \,, \tag{9}$$

where $f_c(p)$ is given by formula (6), and f_i is the density of the i-th marginal distribution of the posterior. Now we show that minimizing the sup-norm of the risk vector is equivalent to equalizing risk vector components.

Lemma 1 *Let $\psi : \Delta^N \to (0, \infty)$ be a continuous nowhere vanishing prior density, let $c \in \mathbb{N}_0^N$ be a count vector from an observation, and let $r(x)$ denote the risk vector arising from integrating the product of a Heaviside loss vector and the posterior density f_c. Then the condition*

$$\|r(\hat{p})\|_\infty = \min_{x \in \Delta^N} \|r(x)\|_\infty = \min_{x \in \Delta^N} \max_{i \in \{1, \ldots, N\}} r_i(x_i) \tag{10}$$

uniquely determines a probability vector $\hat{p} = \hat{p}(c) \in \Delta^N$. Moreover, $\hat{p}$ has the "equalizing property"

$$r_1(\hat{p}_1) = \ldots = r_N(\hat{p}_N). \tag{11}$$

Proof: As the integrand in (9) is continuous and strictly positive, each map r_i is continuous and strictly decreasing on $[0, 1]$ from $r_i(0) = 1$ to $r_i(1) = 0$. Consequently, the inverse functions $r_i^{-1} : [0, 1] \to [0, 1]$ are also continuous and strictly decreasing. Therefore, the function

$$S : [0, 1] \to [0, N], \quad S(\varrho) := \sum_{i=1}^{N} r_i^{-1}(\varrho)$$

is also continuous and strictly decreasing from $S(0) = N$ to $S(1) = 0$. Now the intermediate value theorem gives a value $\varrho_0 \in [0, 1]$ such that $S(\varrho_0) = 1$, and from strict monotonicity of S we infer that ϱ_0 is uniquely determined. Hence, there is a unique vector $\hat{x} \in \Delta^N$ sharing the equalizing risks property (11).

Next we show that $\hat{x}$ also minimizes the maximum of risk vector components. For doing this, assume that there is another probability vector $y \in \Delta^N$ with

$$\max_{i\in\{1,\ldots,N\}} r_i(y_i) < \max_{i\in\{1,\ldots,N\}} r_i(\hat{x}_i) = \varrho_0.$$

This is only possible if, for each index $1 \leq i \leq N$, we have $r_i(y_i) < \varrho_0 = r_i(\hat{x}_i)$. Then strict monotonicity of the r_i^{-1} implies

$$\sum_{i=1}^{N} y_i > \sum_{i=1}^{N} \hat{x}_i = 1,$$

contradicting $y \in \Delta^N$. ∎

Now we take $\hat{p}(c)$ as an estimator for the unknown probability vector $p \in \Delta^N$ governing the process leading to the observations, and we consider the problem of consistency. The following result is fundamental.

Lemma 2 *Let $c(n) = (c_1(n), \ldots, c_N(n))$ be a sequence of count vectors, and put $c_0(n) := \sum c_i(n)$. If, for some index $i_0 \in \{1, \ldots, N\}$, the relative frequencies $c_{i_0}(n)/c_0(n)$ converge to some $\mu_{i_0} \in [0, 1]$, then also $\hat{p}_{i_0}(c(n))$ converges to μ_{i_0}.*

Proof: Suppose that there is a subsequence $(c(n_k))_{k\in\mathbb{N}}$ such that

$$\lim_{k\to\infty} \hat{p}_{i_0}(c(n_k)) = \mu'_{i_0} > \mu_{i_0}. \tag{12}$$

W.l.o.g., we can assume that this subsequence has the property that $\hat{p}_i(c(n_k))$ has a limit μ_i for each $i \in \{1, \ldots, N\}$. As for each fixed k, we have

$$\sum_i \hat{p}_i(c(n_k)) = 1,$$

there must exist an index $j \in \{1, \ldots, N\} \setminus \{i_0\}$ with

$$\lim_{k\to\infty} \hat{p}_j(n_k) = \mu'_j < \mu_j. \tag{13}$$

Denoting by $f_{c(n),i}(\xi)$ the i-th marginal density of the posterior after $c_0(n)$ observations, we get from (11) the equation

$$\int_{\hat{p}_i(c(n))}^{1} f_{c(n),i}(\xi)\, d\xi = \int_{\hat{p}_j(c(n))}^{1} f_{c(n),j}(\xi)\, d\xi\,. \tag{14}$$

Now recall the (since Laplace) well-known fact that the i-th marginal distribution of the posterior converges to the one-point distribution concentrated in μ_i. Hence, (12) implies that the left hand side of (14) converges to 0, and (13) implies that the right hand side of (14) converges to 1, contradicting equality. ∎

In order to state what is meant by consistency, fix a probability vector $p \in \Delta^N$, and assume that we have an infinite sequence of observations. Let

$$\mathcal{S} := \{E_1, \ldots, E_N\}^{\mathbb{N}}$$

denote the set of possible outcome sequences. To each sequence of outcomes $E = (E(n))_{n\in\mathbb{N}} \in \mathcal{S}$ we assign the sequence of count vectors $c(E, n)$ with components

$$c_i(E, n) := |\{k \in \{1, \ldots, n\} : E(k) = E_i\}|.$$

We consider two types of consistency.

1. Suppose that p governs a multinomial process, and denote by P the probability measure on $\mathcal{S}$ induced by p. Then $\hat{p}$ is a *frequentist consistent* estimator for p, if, for any real $a > 0$,

$$\lim_{n\to\infty} P(\|\hat{p}(c(E,n)) - p\| \geq a) = 0. \tag{15}$$

2. $\hat{p}$ is called a *Bayesian consistent* estimator for p, if, for any concrete sequence $(c(n))_{n\in\mathbb{N}}$ of count vectors satisfying

$$\frac{c(n)}{\sum_{i=1}^N c_i(n)} \quad \xrightarrow{n\to\infty} \quad p, \tag{16}$$

we have both $\hat{p}(c(n)) \to p$, and the sequence of posteriors $\Pi_{c(n)}$ converges in distribution to the one-point distribution concentrated in p.

Theorem 3 *Let $p \in \Delta^N$. Then $\hat{p}$ defined by* (11) *is an estimator for p which is both frequentist consistent and Bayesian consistent.*

Proof: In order to prove frequentist consistency, observe that the strong law of large numbers implies that

$$\frac{c(E,n)}{\sum_{i=1}^N c_i(E,n)} \xrightarrow{n\to\infty} p \qquad \text{almost surely.}$$

We infer from lemma 2 that

$$\lim_{n\to\infty} \hat{p}(c(E,n)) = p \quad \text{almost surely,}$$

which implies (15).

In order to see Bayesian consistency, let $(c(n))_{n\in\mathbb{N}}$ satisfy the convergence condition (16). Then lemma 2 proves

$$\lim_{n\to\infty} \hat{p}(c(n)) = p.$$

Convergence of the posteriors $\Pi_{c(n)}$, which are given by their densities (6), to the one-point-distribution concentrated in p, is again the result of Laplace already mentioned in the proof of lemma 2. ∎

5 The Equalizing-Risks Algorithm

We will now use the well-known fact that, if the prior is a Dirichlet distribution, then the posterior is again a Dirichlet distribution, with parameters adjusted using the observation. More precisely, let

$$a = (a_1, \ldots, a_N) \in \mathbb{N}^N$$

be a multi-index. Then the Dirichlet distribution $\mathrm{Dir}(a)$ has density

$$f(x;a) = \frac{x^{a-1}}{B(a)} := \frac{1}{B(a)} \prod_{j=1}^N x_i^{a_i-1},$$

where the normalization constant is given by $B(a) = \int_{\Delta^N} x^{a-1} dS(x)$. If we make an observation with a count vector $c = (c_1, \ldots, c_N)$, then the posterior is the Dirichlet distribution with parameters

$$b = (b_1, \ldots, b_N) := (a_1 + c_1, \ldots, a_N + c_N).$$

With the notations $a_0 = \sum_{i=1}^N a_i$ and $b_0 = \sum_{i=1}^N b_i$, the i-th marginal distribution is a beta distribution with parameters

$$(b_i, b_0 - b_i) = (a_i + c_i, a_0 + c_0 - a_i - c_i). \tag{17}$$

It follows that the marginal density f_i is given by

$$f_i(\xi) = \frac{\Gamma(b_0)}{\Gamma(b_i)\Gamma(b_0 - b_i)} \xi^{b_i-1}(1-\xi)^{b_0-b_i-1} \qquad \text{for } 0 \leq \xi \leq 1.$$

In the case of the uniform prior density $\psi(p) \equiv 1$, the prior is the Dirichlet distribution with parameters $a = (1, \ldots, 1)$. In this case, we obtain $a_0 = N$, and the formula

$$f_i(\xi) = \frac{\Gamma(c_0 + N)}{\Gamma(c_i + 1)\Gamma(c_0 + N - c_i - 1)} \, \xi^{c_i} (1 - \xi)^{c_0 + N - c_i - 2}. \tag{18}$$

In the following we will restrict our considerations to this particular case of a uniform prior. The reader should be easily able, however, to adapt this case to the more general of a prior having a Dirichlet distribution with $a \neq (1, \ldots, 1)$.

We further recall the properties of risk vectors (9). Each component

$$r_i(x_i) = \int_{x_i}^{1} f_i(\xi) \, d\xi, \tag{19}$$

where f_i is according to (18), is a continuous and strictly monotonic function of x_i with values decreasing from 1 to 0. Then, all components of the vector valued function v, where

$$v : [0, 1] \ni A \mapsto (v_1(A), \ldots, v_N(A)) \in \mathbb{R}^N, \quad v_i(x) := r^{-1}(x) \tag{20}$$

are continuous and strictly decreasing. In order to perform minimization (10), we have to equalize risks, i. e., we have to choose $\hat{p}$ such that

$$r_i(\hat{p}) = A_0 \quad \text{for } i \in \{1, \ldots, N\} \tag{21}$$

for some constant $A_0 > 0$, which is, as we have seen in the proof of Lemma 1, Sect. 4, uniquely determined by the constraint $\sum \hat{p}_i = 1$. For finding A_0, we have to solve the equation

$$S(A) = 1, \text{ where } S(A) := \sum_{i=1}^{N} v_i(A).$$

In order to determine $v_i(A)$ as dependent on A, we solve the equations

$$\int_{\xi^{(i)}}^{1} f_i(x) dx = A \tag{22}$$

for $\xi^{(i)} = v_i(A)$. Approximately, the solutions can be found by Newton's method. In applying this method, the corresponding recursion is

$$\xi_{n+1}^{(i)} = \xi_n^{(i)} + \frac{\int_{\xi_n^{(i)}}^{1} f_i(x) dx - A}{f_i(\xi^{(i)})}.$$

The problem is that the denominator $f_i(\xi^{(i)})$ and its derivative (both quantities are crucial for the convergence properties of Newton's method) may attain very large values, which fact makes certain modifications necessary: We introduce the substitution

$$\xi^{(i)} = \widetilde{\mu}_i + \alpha^{(i)} \widetilde{\sigma}_i,$$

where

$$\widetilde{\mu}_i := \frac{c_i}{c_0 + N - 2}, \qquad \widetilde{\sigma}_i := \sqrt{\frac{\widetilde{\mu}_i (1 - \widetilde{\mu}_i)}{c_0 + N - 2}}.$$

Instead of solving equation (22) for $\xi^{(i)}$, we solve

$$\int_{\widetilde{\mu}_i + \alpha^{(i)} \widetilde{\sigma}_i}^{1} f_i(x) dx = A$$

for $\alpha^{(i)}$. Observing that the substitution $x = \widetilde{\mu}_i + z \widetilde{\sigma}_i$ gives

$$\int_{\widetilde{\mu}_i + \alpha^{(i)} \widetilde{\sigma}_i}^{1} f_i(x) dx = \int_{\alpha^{(i)}}^{(1 - \widetilde{\mu}_i)/\widetilde{\sigma}_i} g_i(z) dz,$$

where

$$g_i(z) := \frac{\Gamma(c_0+N)}{\Gamma(c_i+1)\Gamma(c_0+N-c_i-1)}\widetilde{\mu}_i^{c_i}(1-\widetilde{\mu}_i)^{c_0+N-c_i-2}\widetilde{\sigma}_i \times \\ \times \left(1+z\frac{\widetilde{\sigma}_i}{\widetilde{\mu}_i}\right)^{c_i}\left(1-z\frac{\widetilde{\sigma}_i}{1-\widetilde{\mu}_i}\right)^{c_0+N-c_i-2}$$

is bounded from above by 1, we obtain the recursion

$$\alpha_{n+1}^{(i)} = \alpha_n^{(i)} + \frac{\int_{\alpha_n^{(i)}}^{(1-\widetilde{\mu}_i)/\widetilde{\sigma}_i} g_i(z)dz - A}{g_i(\alpha_n^{(i)})}.$$

For $c_i \neq 0$ with growing x the graph of the function

$$\alpha \mapsto \int_{\alpha}^{(1-\widetilde{\mu}_i)/\widetilde{\sigma}_i} g_i(z)dz \tag{23}$$

changes from concavity to convexity at the inflection point $z = 0$. Therefore, if we start with $\alpha_0^{(i)} = 0$, the procedure converges in any case. For $c_i = 0$ the procedure converges as well starting from $\alpha_0^{(i)} = 0$, because of the thorough concavity of (23) in this case.

In order to determine A_0 we have to solve the equation

$$\sum_{i=1}^{N} v_i(A) = 1$$

for A. In principle, this could be done by Newton's method as well. For the sake of numerical robustness, however, we recommend to use the bisection method in this situation.

For dealing with constraints (1) as described in the introduction, let m be a fixed positive real number satisfying $Nm < 1$, and define the set of all probability vectors satisfying the constraints,

$$\Delta' := \left\{x \in \Delta^N : x_i \geq m \text{ for each } i \in \{1,\dots,N\}\right\}.$$

Now we are looking for a probability vector $\hat{p} \in \Delta'$ minimizing the sup-norm of the risk vector,

$$\|r(\hat{p})\|_\infty = \min_{x\in\Delta'} \|r(x)\|_\infty.$$

This problem can be solved by a procedure based on the algorithm sketched above. For modifying our algorithm concerning the function v according to (20), we consider the function

$$w : [0,1] \to \mathbb{R}^N, \quad w_i(A) := \max\{m, v_i(A)\}.$$

The components $A \mapsto w_i(A)$ are still decreasing functions, but no longer strictly decreasing. If we recall the functions ϱ_i defined in (19), we see that w_i is strictly decreasing on $[0, \varrho_i(m)]$ and constant $w_i(t) \equiv m$ for $t \in [\varrho_i(m), 1]$. For arbitrary $A \in [0,1]$, we have the following estimate

$$Nm \leq \sum_{i=1}^{N} w_i(A) \leq N.$$

By continuity of w and the assumption $Nm < 1$, we conclude that there exists $A_0 \in [0,1]$ such that $w(A_0)$ is a probability vector. Moreover, the assumption $Nm < 1$ gives us that $A_0 < r_i(m)$ for at least one index i, which means that $w_i(A_0) > m$ for at least one index i. As each w_i is strictly decreasing on $[0, \varrho_i(m)]$ and gives a bijection $[0, \varrho_i(m)] \to [m, 1]$, this implies that A_0 and hence $w(A_0)$ are uniquely determined. As before, we can evaluate the w_i using Newton's method, and find A_0 by binary search.

6 Equalizing Risks and Asymptotic Statistics

The implementation of the procedure described in the preceding section requires considerable calculational effort, in particular regarding numerical integration. Therefore, the respective computing times are rather long in comparison with "ordinary" discounting methods. In the present section we are going to discuss ideas related to equalizing risks from an asymptotic point of view, which, in the subsequent section, will lead us to a procedure in which equalizing risks is implemented in very good approximation by a modification of square root discounting. The basis of all that follows is a theorem due to Richard von Mises [6] (see appendix A), which shows uniform convergence of the appropriately rescaled posteriors to a multivariate normal distribution. The following theorem can be deduced from this limit theorem as a corollary:

Theorem 4 *Let $\psi : \Delta^N \to \mathbb{R}^+$ be a strictly positive, continuous probability density. Suppose that for $1 \leq i \leq N$ the positive sequences $c_i(n)$ tend to infinity with n, respectively, such that, for $c_0(n) := \sum_{i=1}^N c_i(n)$, the positive limits*

$$p_i := \lim_{n\to\infty} \frac{c_i(n)}{c_0(n)} > 0$$

exist. Let $u_n : \mathbb{R}^N \to \mathbb{R}^+$ be defined by

$$u_n(x) := \begin{cases} C_n \psi(x) \prod_{i=1}^N x_i^{c_i(n)} & \text{if } x \in \Delta^N, \\ 0 & \text{otherwise,} \end{cases}$$

where C_n is the norming constant ensuring u_n being a probability density on Δ^N. Let f_{kn} be the density of the k-th marginal distribution of the distribution assigned to u_n. Then, with the abbreviations

$$a_k(n) := \frac{c_k(n)}{c_0(n)} \quad \text{and} \quad r_k(n) := \sqrt{\frac{a_k(n)(1-a_k(n))}{c_0(n)}}, \tag{24}$$

we have

$$r_k(n) \int_{-\infty}^{t} f_{kn}(a_k(n) + r_k(n)z)dz \xrightarrow{n\to\infty} \Phi(t) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{x^2}{2}} dx$$

for all real t.

Proof: Let $1 \leq s \leq N$ and $s \neq k$. Then, for $z_i = +\infty$ if $i \neq k$ and $i \neq s$, the right side of (31) becomes equal to

$$\frac{1}{\sqrt{2\pi p_k(1-p_k)}} \int_{-\infty}^{z_k} e^{-\frac{x^2}{2p_k(1-p_k)}} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_k/\sqrt{p_k(1-p_k)}} e^{-\frac{x^2}{2}} dx,$$

as can be shown by a straightforward, but cumbersome, calculation. Therefore, (31) implies the limit relation

$$\frac{1}{\sqrt{c_0(n)}} \int_{-\infty}^{z_k} f_{kn}\left(a_k(n) + \frac{z}{\sqrt{c_0(n)}}\right) dz \xrightarrow{n\to\infty} \Phi\left(\frac{z_k}{\sqrt{p_k(1-p_k)}}\right).$$

With the transformation of variables $z = z'\sqrt{a_k(n)(1-a_k(n))}$, we obtain

$$r_k(n) \int_{-\infty}^{z_k/\sqrt{a_k(n)(1-a_k(n))}} f_{kn}(a_k(n) + r_k(n)z')dz' \xrightarrow{n\to\infty} \Phi\left(\frac{z_k}{\sqrt{p_k(1-p_k)}}\right),$$

and, finally, by putting $t := z_k/\sqrt{p_k(1-p_k)}$,

$$U_{kn}\left(\rho_k(n)t\right) \xrightarrow{n\to\infty} \Phi(t),$$

where

$$\rho_k(n) := \frac{\sqrt{p_k(1-p_k)}}{\sqrt{a_k(n)(1-a_k(n))}}, \quad U_{kn}(t) := r_k(n) \int_{-\infty}^{t} f_{kn}(a_k(n) + r_k(n)z')dz'.$$

The convergence of $U_{kn}\left(\rho_k(n)t\right)$ to $\Phi(t)$ being uniform, we get for any real t:

$$\begin{aligned}|U_{kn}(t)-\Phi(t)| &\\ &\leq \left|U_{kn}\left(\rho_k(n)\cdot\frac{t}{\rho_k(n)}\right)-\Phi\left(\frac{t}{\rho_k(n)}\right)\right|+\left|\Phi\left(\frac{t}{\rho_k(n)}\right)-\Phi(t)\right| \\ &\leq \max_{y\in\mathbb{R}}\left|U_{kn}\left(\rho_k(n)y\right)-\Phi(y)\right|+\left|\Phi\left(\frac{t}{\rho_k(n)}\right)-\Phi(t)\right|.\end{aligned}$$

The assertion follows immediately. ∎

We are now able to study asymptotic behavior of our equalizing-risks method, with risk vector components as defined in (9), and compare this to square-root discounting. In order to do this, let us consider a sequence

$$\Big(\big(c_1(n),\ldots,c_N(n)\big)\Big)_{n\in\mathbb{N}}$$

of count vectors related to a sequence of observations such that all conditions of the just stated theorem 4 are met. For configuring our fast discounting algorithm, choose a threshold $m>0$ (satisfying $Nm<1$), initial estimates $\mu_i(n):=a_i(n)$ and $\sigma_i(n):=r_i(n)$, where $a_i(n)$ and $r_i(n)$ are defined in (24). As $r_i(n)$ is proportional to $\sqrt{\mu_i(n)(1-\mu_i(n))}$, this configuration produces exactly the same estimates $\hat{p}_i$ as square-root discounting. Let $\alpha(n)$ be the discounting factor computed by fast discounting. Then, on the basis of theorem 4, for all $1\leq i,j\leq N$ we have

$$\int_{a_i(n)-\alpha(n)r_i(n)}^{1} f_{in}(x)dx-\int_{a_j(n)-\alpha(n)r_j(n)}^{1} f_{jn}(x)dx \xrightarrow{n\to\infty} 0.$$

Therefore, square root discounting "tends" in a certain sense to equalizing risks. It is very notable that this consideration even holds for arbitrary continuous and strictly positive priors ψ.

In "real" situations of application we have a large c_0, the single c_i's may be very small, however. Therefore, in order to adapt the asymptotic idea of an approximate equality of estimates gained by equalizing risks and such obtained by square-root discounting, some modifications of "ordinary" square-root discounting are necessary. In this context, we are only considering uniform priors in the following. Under this assumption, for an individual event E_i, $i=1,\ldots,N$, the posterior distribution for the unknown underlying probability p_i is the i-th marginal of the Dirichlet distribution with parameters

$$(1+c_1,\ldots,1+c_N).$$

According to (17), this is a beta distribution with parameters

$$(a,b)=(c_i+1,c_0+N-c_i-1).$$

This beta distribution has mean

$$\mu_i^{(\beta)}=\frac{a}{a+b}=\frac{c_i+1}{c_0+N}$$

and standard deviation

$$\begin{aligned}\sigma_i^{(\beta)}&=\sqrt{\frac{ab}{(a+b+1)(a+b)^2}}=\frac{1}{c_0+N}\sqrt{\frac{(c_i+1)(c_0+N-c_i+1)}{c_0+N+1}}\\ &=\sqrt{\frac{\mu_i^{(\beta)}\left(1-\mu_i^{(\beta)}\right)}{c_0+N+1}}.\end{aligned}$$

When c_0 tends to infinity, the beta distribution parameters $\mu_i^{(\beta)}$ and $\sigma_i^{(\beta)}$ are asymptotically equivalent to a_i and r_i as defined in (24). Contrary to a_i and r_i, they have the advantage of being always positive. Moreover, from the Bayesian point of view they represent the "natural" estimates for mean and standard deviation of a Bernoulli process. Yet, in order to adapt our asymptotic considerations to the situation, where the parameters $\mu_i^{(\beta)}$ and $\sigma_i^{(\beta)}$ are used instead of the parameters a_i and r_i, we need a corollary of theorem 4:

Corollary 1 *For $\mu_i(n)$ and $\sigma_i(n)$ satisfying the asymptotics*

$$\mu_i(n) \sim a_i(n) \quad \text{and} \quad \sigma_i(n) \sim r_i(n) \qquad \text{for } n \to \infty,$$

the following limit holds:

$$\lim_{n\to\infty} \int_{a_i(n)-\alpha r_i(n)}^{1} f_{in}(x)\,dx = \frac{1}{\sqrt{2\pi}} \int_{-\alpha}^{\infty} \exp\left(-\frac{z^2}{2}\right) dz \quad \forall \alpha > 0.$$

Proof: Set

$$g_{in}(z) = \begin{cases} r_i(n) f_{in}\big(a_i(n) + r_i(n)z\big), & \text{if } a_i(n) + r_i(n)z \in [0,1], \\ 0, & \text{otherwise.} \end{cases}$$

Then theorem 4 implies

$$\lim_{n\to\infty} \int_{-\alpha}^{\infty} g_{in}(z)\,dz = \frac{1}{\sqrt{2\pi}} \int_{-\alpha}^{\infty} \exp\left(-\frac{z^2}{2}\right) dz \quad \forall \alpha > 0.$$

Using $a_i(n) \le 1$ and $r_i(n) \to 0$, we get

$$|a_i(n) - \mu_i(n)| \to 0 \quad \text{and} \quad |r_i(n) - \sigma_i(n)| \to 0 \qquad \text{for } n \to \infty.$$

Hence

$$\begin{aligned} \lim_{n\to\infty} &\int_{\mu_i(n)-\alpha\sigma_i(n)}^{1} f_{in}(x)\,dx = \\ &= \lim_{n\to\infty} \left(\int_{\mu_i(n)-\alpha\sigma_i(n)}^{a_i(n)-\alpha r_i(n)} f_{in}(x)\,dx + \int_{a_i(n)-\alpha r_i(n)}^{1} f_{in}(x)\,dx \right) \\ &= \lim_{n\to\infty} \int_{-\alpha}^{\infty} g_{in}(z)\,dz\,. \end{aligned}$$

∎

7 The Fast Discounting Algorithm with Adjusted Initial Estimates

In this section, we restrict our considerations to the particular case of uniform prior, and therefore to marginal densities f_i according to (18).

As explained above, the connection between equalizing risks and our discounting algorithm is the approximation

$$\int_{\mu_i-\alpha\sigma_i}^{1} f_i(x)dx \approx \frac{1}{\sqrt{2\pi}} \int_{-\alpha}^{\infty} e^{-\frac{x^2}{2}}\,dx, \tag{25}$$

where $\mu_1, \ldots, \mu_N$ are the initial estimates and $\sigma_1, \ldots, \sigma_N$ are the discounting bases used in the fast discounting algorithm described in section 2, and α is the discounting factor computed by that algorithm. Corollary 1 proves that (25) is asymptotically valid whenever $\mu_i \sim a_i$ and $\sigma_i \sim r_i$.

This approximation is rather good if the μ_i are not too small and not too big (that is, too close to 1), but it is not valid with a sufficient goodness of approximation in each case. For very small counts (which in typical situations of application occur quite frequently), or very large counts, such a good approximation is not valid. The basic idea for mastering this problem is to run discounting with adjusted initial estimates but unchanged discounting bases. In order to get appropriate adjustments, we use adjusted initial estimates

$$\mu_i^{(1)} = \mu_i^{(\beta)} + \delta(c_i)\sigma_i \tag{26}$$

such that (25) with μ_i substituted by $\mu_i^{(1)}$ holds with sufficient precision even in cases of small and large counts.

In order to find those $\delta(c_i)$, we first determine a "provisional" discounting factor $\alpha^{(0)}$ by use of the discounting algorithm with initial estimates $\mu_i^{(0)} := \mu_i^{(\beta)}$ and discounting bases $\sigma_i := \sigma_i^{(\beta)}$. Next, we calculate $\delta(c_i)$ for small and large c_i through the condition

$$\int_{\mu_i^{(1)}-\alpha^{(0)}\sigma_i}^{1} f_i(x)dx = \frac{1}{\sqrt{2\pi}} \int_{-\alpha^{(0)}}^{\infty} e^{-\frac{x^2}{2}}\,dx,$$

where f_i denotes the beta density assigned to c_i. As we will see in the following, for sufficiently great c_0 ($c_0 \geq 10^5$, say) and small (as well as large) counts c_i ($c_i \leq 1000$ and $c_i \geq c_0 - 1000$), the shift constants $\delta(c_i)$ depend only on c_i and $\alpha^{(0)}$ with good precision, such that they may be computed in advance and stored in an appropriate buffer. Finally, we run the discounting algorithm with initial estimates $\mu_i^{(1)}$ according to (26) ($\delta(c_i)$ being equal to zero for counts c_i which are neither very small nor very large) and discounting bases σ_i. If the resulting discounting factor $\alpha^{(1)}$ does not differ too much from $\alpha^{(0)}$, then we can be sure that the thereby gained estimates $\hat{p}_i$ are approximately equal to those obtained by equalizing risks.

At least in important cases of application, an explicit discussion of the mutual closeness of $\alpha^{(0)}$ and $\alpha^{(1)}$ is possible: With $c_k \in \mathbb{N}_0$ we consider the count vector $(c_1, c_2, \ldots, c_N)$ of length N, and, as usual, we use the abbreviation $c_0 = \sum_{k=1}^N c_k$. We further presuppose a fixed $M > 0$ such that the constant m meets the following equation:

$$m = \frac{M}{c_0 + N}.$$

Finally, by μ we denote the maximum of all "provisional" initial estimates $\mu_i^{(0)} = \mu_i^{(\beta)}$, and we assume $\mu < 1/2$. The latter assumption refers to an important class of applications. In principle, the methods employed are applicable to more general situations, but a comparably thorough treatment would require considerably more effort.

In a first step, we discuss the discounting algorithm with initial estimates $\mu_i^{(0)}$ and discounting bases σ_i. If $M \leq 1$, then the algorithm terminates already after the first step, and we obtain for the solution α the value $\alpha^{(0)} = 0$. If $M > 1$, and if a solution α is only reached after some repetitions, then there is a certain index $J \in \{1, \ldots, N\}$ in the recursion part of the algorithm with the property $\alpha_{J-1} < \alpha \leq \alpha_J$. Putting $r_1 := J - 1$, we get

$$\frac{\sum_{k=r_1}^N \mu_k^{(0)} - 1 + (r_1 - 1)m}{\sum_{k=r_1}^N \sigma_k} > \frac{\mu_{r_1}^{(0)} - m}{\sigma_{r_1}}$$

and

$$\alpha^{(0)} := \frac{\sum_{k=r_1+1}^N \mu_k^{(0)} - 1 + r_1 m}{\sum_{k=r_1+1}^N \sigma_k} \leq \frac{\mu_{r_1+1}^{(0)} - m}{\sigma_{r_1+1}}.$$

Moreover, as $\sum_{i=1}^N \mu_i^{(0)} = 1$, it is obvious that $\alpha^{(0)} > 0$.

Using the generalized triangle inequality and taking into consideration

$$\sum_{k=r_1+1}^N \mu_k^{(0)} = 1 - r_1 m + \alpha^{(0)} \sum_{k=r_1+1}^N \sigma_k,$$

we obtain

$$\begin{aligned}
\sum_{k=r_1+1}^N \sigma_k &= \frac{1}{\sqrt{c_0 + N + 1}} \sum_{k=r_1+1}^N \sqrt{\mu_k^{(0)}(1 - \mu_k^{(0)})} \\
&\geq \frac{1}{\sqrt{c_0 + N + 1}} \sqrt{\sum_{k=r_1+1}^N \mu_k^{(0)}(1 - \mu_k^{(0)})} \\
&\geq \frac{1}{\sqrt{c_0 + N + 1}} \sqrt{(1 - \mu) \sum_{k=r_1+1}^N \mu_k^{(0)}} \\
&\geq \frac{\sqrt{(1 - \mu)(1 - r_1 m)}}{\sqrt{c_0 + N + 1}},
\end{aligned}$$

where we also used that $\mu_k^{(0)} \leq \mu$ for $k \geq r_1$. Finally, we get an upper bound

$$\begin{aligned}\alpha^{(0)} &\leq \left(\sum_{k=1}^{N} \mu_k^{(0)} - 1 + r_1 m\right) \frac{\sqrt{c_0+N+1}}{\sqrt{(1-\mu)(1-r_1 m)}} \\ &= r_1 m \frac{\sqrt{c_0+N+1}}{\sqrt{(1-\mu)(1-r_1 m)}} \\ &\leq (N-1)m \frac{\sqrt{c_0+N+1}}{\sqrt{(1-\mu)(1-(N-1)m)}} \\ &= \frac{(N-1)M}{\sqrt{(1-\mu)(c_0+N-(N-1)M)}} \sqrt{1+\frac{1}{c_0+N}}. \end{aligned} \tag{27}$$

On the other hand, a lower bound for $\mu_i^{(0)}$ can be obtained by

$$\frac{\mu_i^{(0)} - m}{\sigma_i} = \frac{c_i + 1 - M}{\sqrt{\frac{(c_i+1)(c_0+N-c_i-1)}{c_0+N+1}}} > \sqrt{c_i+1} - \frac{M}{\sqrt{c_i+1}}. \tag{28}$$

We are now heading to an estimate for $\alpha^{(0)}$ under realistic assumptions. From (27) we can see that, under the realistic assumption of a very large c_0, the upper bound of $\alpha^{(0)}$ is rather small. Thus, it is reasonable to make the modest assumption that this upper bound is not exceeding 1. On the basis of (28), we conclude that the maximum index r_1 at which the algorithm terminates has the property

$$1 > \sqrt{c_{r_1+1}+1} - \frac{M}{\sqrt{c_{r_1+1}+1}},$$

or, equivalently,

$$c_{r_1+1} > \frac{2M - 1 + \sqrt{1+4M}}{2} \geq \frac{1+\sqrt{1+4M}}{2}.$$

Thus, under the realistic condition that, for large c_0, by far the most of the counts are above this rather small bound, a rough estimate can be obtained using the approximations

$$\sum_{i=r_1+1}^{N} \mu_i^{(0)} \approx \sum_{i=1}^{N} \mu_i^{(0)} \quad \text{and} \quad \sum_{i=r_1+1}^{N} \sigma_i \approx \sum_{i=1}^{N} \sigma_i.$$

For our next inequality, we apply lemma 7 from appendix B. We arrange the counts c_i such that $c_1 \leq \ldots \leq c_N$, and put

$$\mu' := \frac{c_N}{c_0}.$$

Then

$$\left[\frac{1}{\mu'}\right] c_N = \left[\frac{1}{\mu'}\right] \mu' c_0 \leq c_0 = \sum_{i=1}^{N} c_i,$$

and, by lemma 7 with $z_i = c_i$ and $f(z) = \sqrt{z}$, we infer

$$\sum_{i=1}^{N} \sqrt{c_i} \geq \left[\frac{1}{\mu'}\right] \sqrt{\mu' c_0}.$$

Presupposing $N \ll c_0$, we obtain $\mu' \approx \mu$ and

$$\frac{\sqrt{1-\mu}}{\sqrt{\mu c_0}} \lessapprox \sum_{i=r_1+1}^{N} \sigma_i \approx \sum_{i=1}^{N} \sigma_i.$$

Therefrom, it follows an approximate upper bound for $\alpha^{(0)}$ according to

$$\alpha^{(0)} \approx \frac{r_1 m}{\sum_{i=1}^{N} \sigma_i} \lessapprox \frac{r_1 M \sqrt{\mu}}{\sqrt{c_0(1-\mu)}}. \tag{29}$$

These considerations evidence that in many realistic situations $\alpha^{(0)}$ will be rather close to zero. Therefore, in the following we will focus on the assumption $\alpha^{(0)} = 0$.

We recall the well known fact that, for $c_i \ll c_0$, a very good approximation to the standardized beta density

$$g_i(x) = \begin{cases} k(c_i, c_0, N) f(\mu_i^{(0)} + \sigma_i x) & \text{for} \quad -\frac{\mu_i^{(0)}}{\sigma_i} \le x \le \frac{1-\mu_i^{(0)}}{\sigma_i} \\ 0 & \text{else,} \end{cases}$$

$k(c_i, c_0, N)$ being the norming constant, is provided by

$$\widetilde{g}_i(x) := \frac{(c_i+1)^{(2c_i+1)/2}}{c_i!\, e^{c_i+1}} \left(1 + \frac{x}{\sqrt{c_i+1}}\right)^{c_i} e^{-\sqrt{c_i+1}\, x}.$$

Independent of c_0, we are therefore able to improve the quality of approximation in (25) for small counts by adjusting the initial estimates μ_i by taking

$$\mu_i^{(1)} := \mu_i^{(0)} + \delta(c_i)\sigma_i,$$

where $\delta(c_i)$ is defined by the condition

$$\int_{\mu_i^{(0)} + \delta(c_i)\sigma_i}^{1} f_i(x)dx = \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{-\frac{x^2}{2}} dx = \frac{1}{2}.$$

Because of

$$\int_{\mu_i^{(1)}}^{1} f_i(x)dx = \int_{\delta(c_i)}^{\infty} g_i(x)dx \approx \int_{\delta(c_i)}^{\infty} \tilde{g}_i(x)dx,$$

we finally reach the condition

$$\int_{\delta(c_i)}^{\infty} \tilde{g}_i(x)dx = \frac{1}{2}.$$

A numerical evaluation (∞ being replaced by numbers such that the neglected part contributes to the whole integral with an amount less than 10^{-14}) yields monotonically increasing negative $\delta(c_i)$, where $\delta(0)$ to $\delta(11)$ are (with an error about $\pm 10^{-8}$) equal to

-0.30685281	-0.22744302	-0.18818136	-0.16396962
-0.14717402	-0.13465613	-0.12486544	-0.11693798
-0.11034960	-0.10476163	-0.09994425	-0.09573531

Already from a relatively small c_i on, the constants $\delta(c_i)$ are changing only slowly. We have $\delta(100) = -0.03314837$, $\delta(500) = -0.01489047$, and $\delta(1000) = -0.01053503$. The relative deviation

$$\frac{\mu_i^{(0)} - \mu_i^{(1)}}{\mu_i^{(0)}}$$

is less than about 0.003 from $c_i = 100$ on. On the other hand, only with counts from about 1000 on the relative error in integrating without the correction by the δ's becomes less than 10^{-2}. Therefore, we substitute $\mu_i^{(0)}$ by $\mu_i^{(1)}$ for $c_i \le 999$, and we set $\delta(c_i) = 0$ for $i \ge 1000$.

Now, the important question is: If we apply our discounting algorithm with $\mu_i^{(1)}$ instead of $\mu_i^{(0)}$, what influence will this have on the behavior of the "new" $\alpha^{(1)}$ which replaces the "old" $\alpha^{(0)}$? First, the application of the shifting constants $\delta(c_i)$ does not change the order of the α_i. In fact, we have

$$c_i \le c_{i+1} \Leftrightarrow \frac{\mu_i^{(0)} - m}{\sigma_i} \le \frac{\mu_{i+1}^{(0)} - m}{\sigma_{i+1}} \Leftrightarrow \frac{\mu_i^{(1)} - m}{\sigma_i} \le \frac{\mu_{i+1}^{(1)} - m}{\sigma_{i+1}},$$

the first relation being valid due to the monotonicity of the function

$$]0;1[\ni x \mapsto \frac{x-m}{\sqrt{x(1-x)}} \quad \left(0 \le m < \frac{1}{N}\right),$$

the second due to the monotonicity of $\delta(c_i)$. Second, likewise due to the latter property, the algorithm does not terminate earlier, because

$$\frac{\sum_{i=r}^N \mu_i^{(0)} - 1 - (r-1)m}{\sum_{i=r}^N \sigma_i} > \frac{\mu_r^{(0)} - m}{\sigma_r} \Rightarrow \frac{\sum_{i=r}^N \mu_i^{(1)} - 1 - (r-1)m}{\sum_{i=r}^N \sigma_i} > \frac{\mu_r^{(1)} - m}{\sigma_r}.$$

Third, if only

$$-\delta(c_{r+1}) \le \frac{\mu^{(0)}_{r_1+1} - m}{\sigma_{r_1+1}},$$

then the algorithm with the adjusted $\mu_i^{(1)}$ terminates at the same index r_1 as the algorithm with $\mu_i^{(0)}$. In this case, $\alpha^{(1)}$ is less than $\alpha^{(0)}$ (in any case, $\alpha^{(1)}$ is greater than $\delta(c_1)$). And fourth, since the $\alpha^{(1)}$'s are always less or equal than the corresponding $\alpha^{(0)}$, the approximate upper bound (29) is also valid for $\alpha^{(1)}$ under the same conditions.

In order to obtain an approximate lower bound for $\alpha^{(1)}$, a little more effort is necessary. First we observe that, due to the general relation

$$\frac{a}{b} > \frac{c}{d} \Leftrightarrow \frac{a-c}{b-d} > \frac{a}{b} \quad (b > d > 0),$$

for $r \ge 1$ the inequality

$$\frac{\sum_{i=r}^{N} \mu_i^{(1)} - 1 + (r-1)m}{\sum_{i=r}^{N} \sigma_i} > \frac{\mu_r^{(1)} - m}{\sigma_r}$$

implies

$$\frac{\sum_{i=r+1}^{N} \mu_i^{(1)} - 1 + rm}{\sum_{i=r+1}^{N} \sigma_i} > \frac{\sum_{i=r}^{N} \mu_i^{(1)} - 1 + (r-1)m}{\sum_{i=r}^{N} \sigma_i}.$$

Taking into account the principles of our algorithm we can follow that

$$\alpha^{(1)} \ge \frac{\sum_{i=1}^{N} \mu_i^{(1)} - 1}{\sum_{i=1}^{N} \sigma_i} = \frac{\sum_{i=1}^{N} \delta(c_i)\sigma_i}{\sum_{i=1}^{N} \sigma_i}.$$

Presupposing $N \ll c_0$, and using the approximate lower bound $\sqrt{\frac{1-\mu}{\mu c_0}}$ for $\sum_{i=1}^{N} \sigma_i$, we get

$$-\alpha^{(1)} \lessapprox \frac{\sqrt{\mu}\sum_{i=1}^{N}(-\delta(c_i)\sqrt{c_i+1})}{\sqrt{c_0(1-\mu)}}.$$

A numerical evaluation shows that

$$\delta(c_i)\sqrt{c_i+1} \le 0.34.$$

Therefore, if N_1 denotes the number of events with counts less than 1000, then we finally obtain

$$\alpha^{(1)} \ge -\frac{0.34 N_1 \sqrt{\mu}}{\sqrt{c_0(1-\mu)}}.$$

In the characteristic case of $N_1 = 100$ and $c_0 = 10^6$ we thusly have -0.034 as a lower bound of $\alpha^{(1)}$.

8 The Practical Example Revisited

We again consider the count vector given in (5) with length $N = 131$ and $c_0 = 662623$, and start by applying our equalizing algorithm with threshold $m = 10^{-5}$. All calculations are performed by use of the numeric system "Euler" (`http://www.rene-grothmann.de/euler.html`; thanks to its author, René Grothmann, for introducing us into the specific features of this program). The

results are (up to a precision of 10^{-9}):

c_i	$\hat{p}_i$	c_i	$\hat{p}_i$	c_i	$\hat{p}_i$
0	$1 \cdot 10^{-5}$	1	$1 \cdot 10^{-5}$	2	$1 \cdot 10^{-5}$
4	$1 \cdot 10^{-5}$	7	$1.139 \cdot 10^{-5}$	18	$2.7882 \cdot 10^{-5}$
61	$9.2528 \cdot 10^{-5}$	102	0.000154241	104	0.000157252
247	0.000372654	268	0.000404297	278	0.000419366
293	0.000441969	350	0.000527869	426	0.000642415
463	0.000698185	571	0.000860984	571	0.000860984
572	0.000862492	614	0.000925807	779	0.001174562
815	0.001228838	872	0.001314779	928	0.001399213
965	0.001455001	1095	0.001651019	1224	0.001845536
1288	0.001942044	1353	0.002040061	1425	0.002148635
1913	0.002884563	1984	0.002991639	2065	0.003113797
2068	0.003118321	2156	0.003251038	2169	0.003270644
2199	0.003315888	2327	0.003508933	2386	0.003597916
2699	0.004069984	2861	0.004314318	2885	0.004350516
3017	0.004549605	3090	0.004659709	3207	0.004836176
3267	0.004926673	3270	0.004931198	3531	0.005324863
3804	0.005736633	4300	0.006484771	5413	0.008163606
5781	0.008718705	6413	0.009672039	6504	0.009809308
6534	0.009854561	6768	0.01020754	7240	0.010919538
7481	0.011283081	7821	0.011795966	7828	0.011806525
8304	0.012524569	8559	0.012909238	8898	0.013420623
9792	0.014769242	10829	0.016333597	11069	0.016695649
12227	0.01844256	12254	0.018483291	12927	0.019498559
13255	0.019993372	13485	0.020340346	15226	0.022966801
15366	0.023178005	15510	0.023395243	15529	0.023423907
18587	0.028037247	19937	0.030073897	22791	0.034379561
32288	0.048707397	47562	0.071751374	56644	0.085453658
65832	0.099315985	77061	0.116257786		

These values correspond to $A = 0.51747026$. In turn, if one calculates the integrals

$$\int_{\alpha^{(i)}}^{\frac{1-\widetilde{\mu}_i}{\widetilde{\sigma}_i}} g_i(x)dx,$$

where

$$\widetilde{\mu}_i + \alpha^{(i)}\widetilde{\sigma}_i = \hat{p}_i,$$

one obtains values differing from A by an amount less than the maximum error which has to be suspected in the numerical determination of these integrals.

In figure 5 we have plotted the differences between estimates computed by adjusted square root discounting and relative frequencies, and between estimates computed by equalizing risks and relative frequencies, respectively. For small up to medium numbers (with the exception of the smallest counts only), it can be shown that the values estimated by equalizing risks are closer to the related relative frequencies than those values computed by square-root discounting. Taking into account the particular form of the $\mu_i^{(1)}$, we can see that for smaller counts a certain positive amount adds to the relative frequencies in the application of equalizing risks, which leads to somewhat greater estimated values when using adjusted square-root discounting. On the other hand, due to the constraint $\sum \hat{p}_i = 1$, for larger counts the differences between relative frequencies and values estimated from equalizing risks are a little greater than the corresponding differences with square-root discounting.

The errors committed in calculating the $\hat{p}_i$'s may be estimated as follows: Our basic precision is, according to the IEEE-754 standards, about 10^{-16}. The integrals are calculated with an adaptive Runge method. The corresponding rounding errors have been checked by a comparison of these results with those obtained by Simpson's method under application of interval arithmetics. The respective upper limits of these integrals are determined by the demand that possibly neglected "tails" (which can be estimated from above by polygons because of their convexity) contribute to the whole integral with an amount less than 10^{-14}. In this way an absolute maximum error of

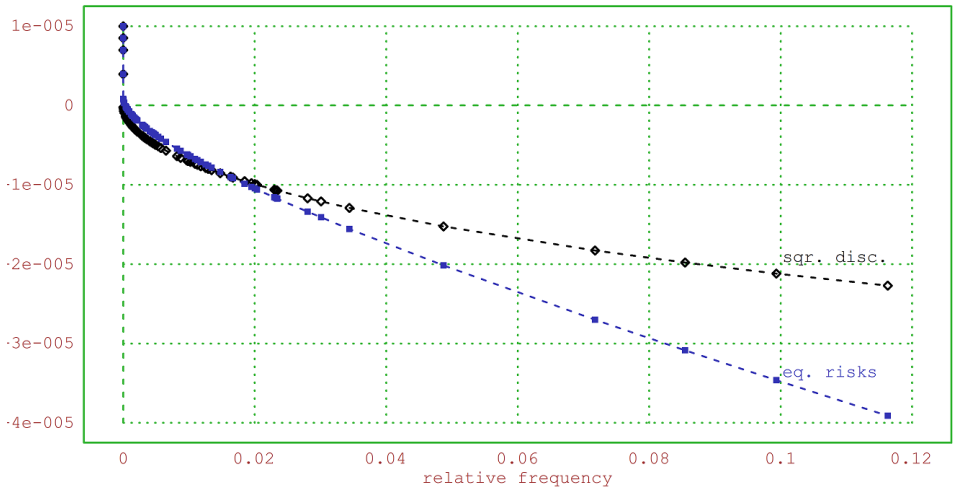

Figure 5: Absolute deviations from relative frequency for estimates obtained by adjusted square-root discounting and equalizing risks, respectively.

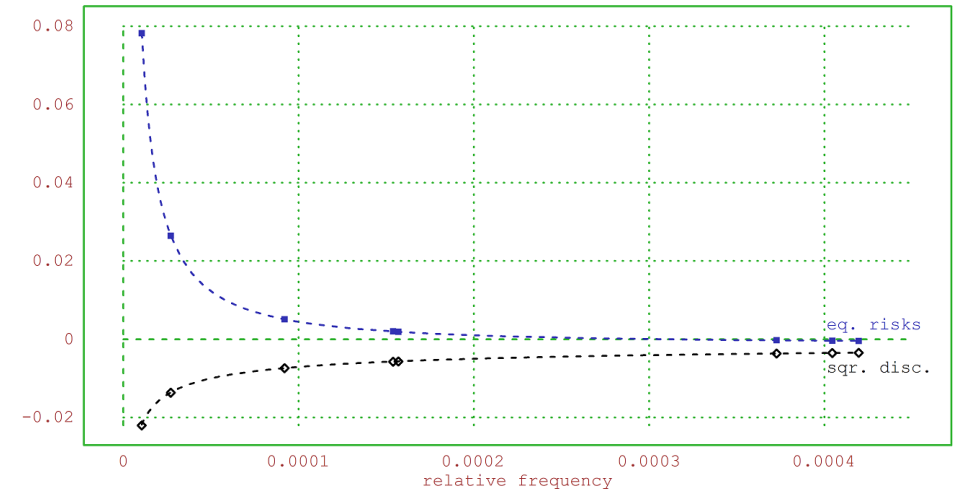

Figure 6: Relative deviations $\frac{\hat{p}_i - q_i}{q_i}$ for square root discounting and equalizing risks, respectively, and relative frequencies q_i; the graphic refers to the counts 7, 18, 61, 102, 104, 247, 268, and 278.

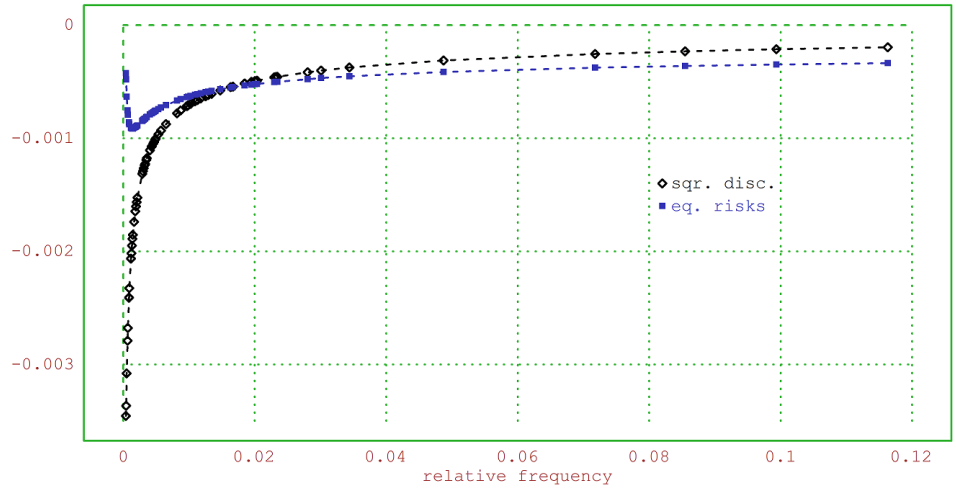

Figure 7: Relative deviations $\frac{\hat{p}_i - q_i}{q_i}$ for square root discounting and equalizing risks, respectively, from relative frequencies q_i for counts ≥ 278.

about $2 \cdot 10^{-11}$ can be guaranteed for the integrals. The norming factors

$$K_i := \frac{\Gamma(c_0+N)}{\Gamma(c_i+1)\Gamma(c_0+N-c_i-1)} \widetilde{\mu}_i^{c_i} (1-\widetilde{\mu}_i)^{c_0+N-c_i-2} \widetilde{\sigma}_i$$

are calculated from the relation

$$K_i^{-1} = \int_{\rho_1}^{\rho_2} \left(1 + x\frac{\widetilde{\sigma}_i}{\widetilde{\mu}_i}\right)^{c_i} \left(1 - x\frac{\widetilde{\sigma}_i}{1-\widetilde{\mu}_i}\right)^{c_0+N-c_i-2} dx,$$

where ρ_1 and ρ_2 are determined such that possibly neglected tails behave as just described.

For a closer determination of the precision of the risk vector $\hat{p}$ we base us on the solution $A \approx 0.52$ for the equation

$$S(A) := \sum_{i=1}^{N} \max(m, v_i(A)) = \sum_{i=1}^{N} \max(m, \widetilde{\mu}_i + \alpha^{(i)}(A)\widetilde{\sigma}_i) = 1 \tag{30}$$

in our case. The iteration for calculating the function values $v_i(A)$ by Newton's method is terminated in each case if the deviation between two successive values is below 10^{-10}. By again applying interval arithmetics it can be shown that the $\alpha^{(i)}(A)$ are thus determined with a maximum rounding error of ca. $1.2 \cdot 10^{-10}$.

For approximately solving (30) the bisection method is used. Neglecting the very small rounding errors of $\widetilde{\mu}_i$ and $\widetilde{\sigma}_i$, we obtain for the rounding error $\delta S(A)$:

$$\delta S(A) \approx \sum_{i=1}^{N} \widetilde{\sigma}_i \delta\alpha^{(i)}(A) \approx 0.09 \cdot 1.2 \cdot 10^{-11} \approx 1.1 \cdot 10^{-11}.$$

The rounding error δA can be derived from the equation

$$\delta S(A) = \sum_{i=1}^{N} \delta w_i(A) \approx \sum_{i=i_0}^{N} v_i'(A)\delta A,$$

where i_0 denotes the maximum index such that $w_i(A) = m$ for $i < i_0$. In our case, $i_0 = 53$, and

$$\sum_{i=i_0}^{N} v_i'(A) = \sum_{i=i_0}^{N} -\frac{\widetilde{\sigma}_i}{g_i(\alpha^{(i)})} \approx -0.02.$$

Therefore we obtain

$$\delta A \approx \frac{1.1 \cdot 10^{-11}}{0.02} = 5.5 \cdot 10^{-10}.$$

The bisection terminates when $S(a) \le 1$ and $S(b) > 1$ for $|a-b| < 10^{-8}$. The rounding errors for a and b of ca. $5.5 \cdot 10^{-10}$ being far less than 10^{-8}, for the true A we therefore have a maximum error ΔA of approximately 10^{-8} as well. Finally, for estimating the error in the calculation of $\hat{p}$ we notice that, because of

$$\int_{\alpha^{(i)}}^{\frac{1-\widetilde{\mu}_i}{\widetilde{\sigma}_i}} g_i(x)dx = A,$$

we have

$$\Delta\alpha^{(i)} g_i(\alpha^{(i)}) \approx \Delta A \quad (i \ge i_0),$$

and therefore

$$\Delta\alpha^{(i)} \approx \frac{\Delta A}{0.4}.$$

This implies

$$\Delta\hat{p}_i \approx \widetilde{\sigma}_i \Delta\alpha^{(i)} < \frac{0.0004\Delta A}{0.4} = 10^{-11}.$$

Applying the square-root discounting algorithm for $m = 10^{-5}$ with the above-described shift corrections, we obtain

$$r_1 = 53 \quad \text{and} \quad \alpha^{(1)} = 0.04711690100541.$$

The results for the $\hat{p}$'s according to adjusted square-root discounting, $\hat{p}_i^{\mathrm{sqa}}$, and equalizing risks, $\hat{p}_i^{\mathrm{er}}$, are exemplified and compared in the following table. Here the precision of all values in the table is $\pm 10^{-9}$, and the relative deviations are computed by the formula

$$\frac{\hat{p}_i^{\mathrm{sqa}} - \hat{p}_i^{\mathrm{er}}}{\hat{p}_i^{\mathrm{er}}}.$$

c_i	$\hat{p}_i^{\mathrm{sqa}}$	$\hat{p}_i^{\mathrm{er}}$	relative dev.
0	$1 \cdot 10^{-5}$	$1 \cdot 10^{-5}$	0
1	$1 \cdot 10^{-5}$	$1 \cdot 10^{-5}$	0
2	$1 \cdot 10^{-5}$	$1 \cdot 10^{-5}$	0
4	$1 \cdot 10^{-5}$	$1 \cdot 10^{-5}$	0
7	$1.1370 \cdot 10^{-5}$	$1.1390 \cdot 10^{-5}$	0.001727421
18	$2.7857 \cdot 10^{-5}$	$2.7882 \cdot 10^{-5}$	0.000922402
61	$9.2486 \cdot 10^{-5}$	$9.2528 \cdot 10^{-5}$	0.000453986
965	0.001454843	0.001455001	0.000108350
1095	0.001651353	0.001651019	−0.000202688
⋮	⋮	⋮	⋮
8898	0.013420643	0.013420623	$-1.507 \cdot 10^{-6}$
9792	0.014769238	0.014769242	$2.53 \cdot 10^{-7}$
⋮	⋮	⋮	⋮
32288	0.048706974	0.048707397	$8.677 \cdot 10^{-6}$
⋮	⋮	⋮	⋮
77061	0.116256867	0.116257786	$7.901 \cdot 10^{-6}$

We observe that, toward larger c_i's, the relative deviation is greater than for medium c_i's where it is negative. This is due to the fact that, for c_i with $c_i c_0 \geq \frac{1}{10}$, there is an almost perfect symmetry of $g_i(x)$ with respect to $x_0 = -\frac{1}{c_0+N}$, but not with respect to $x_0 = 0$ as in case of the normal density. For medium c_i (as, for example, $c_i = 8898$), the slight asymmetry of g_i is very well compensated by this shift of $\frac{1}{c_0+N}$ toward the left.

Appendix A

Von Mises [6, p. 86] essentially proved the following

Theorem 5 *Let $(c_1(n))_{n\in\mathbb{N}}, \ldots, (c_N(n))_{n\in\mathbb{N}}$ be N sequences of positive real numbers, each diverging to ∞, and $c_0(n) := \sum_{i=1}^{N} c_i(n)$; suppose that, for each index $i \in \{1, \ldots, N\}$, the positive limit*

$$p_i = \lim_{n\to\infty} a_i(n), \qquad \textit{where } a_i(n) := \frac{c_i(n)}{c_0(n)}$$

exists. Let s be a fixed index with $1 \leq s \leq N$, and, for

$$z := (z_1, \ldots, z_{s-1}, z_{s+1}, \ldots, z_N) \quad (z_i \in \mathbb{R}),$$

let the quadratic form $Q(z)$ be defined by

$$Q(z) := \frac{1}{2}\left(\sum_{\substack{i=1\\ i\neq s}}^{N} \frac{z_i^2}{p_i} + \frac{1}{p_s}\left(\sum_{\substack{i=1\\ i\neq s}}^{N} z_i\right)^2\right).$$

Finally, with the abbreviation

$$D^N := \left\{ x \in \mathbb{R}^N \,\middle|\, \sum_{i=1}^{N} x_i = 1 \right\},$$

let $\psi : D^N \to \mathbb{R}^+$ be a function which is strictly positive and continuous in Δ^N, and identical to zero in $D^N \setminus \Delta^N$, and, for z as defined above, let

$$\widetilde{\psi}_n(z) := \psi \left(a_1(n) + \frac{z_1}{\sqrt{c_0(n)}}, \ldots, a_{s-1}(n) + \frac{z_{s-1}}{\sqrt{c_0(n)}}, \right.$$
$$\left. a_s - \frac{1}{\sqrt{c_0(n)}} \sum_{\substack{i=1 \\ i \neq s}}^{N} z_i, a_{s+1}(n) + \frac{z_{s+1}}{\sqrt{c_0(n)}}, a_N(n) + \frac{z_N}{\sqrt{c_0(n)}} \right).$$

Then, for z as above, the functions $w_n(z)$, defined by

$$w_n(z) := K_n \widetilde{\psi}_n(z) \left(a_s(n) - \frac{1}{\sqrt{c_0(n)}} \sum_{\substack{i=1 \\ i \neq s}}^{N} z_i \right)^{c_s(n)} \prod_{\substack{i=1 \\ i \neq s}}^{N} \left(a_i(n) + \frac{z_i}{\sqrt{c_0(n)}} \right)^{c_i(n)},$$

where K_n is the norming constant ensuring

$$\int_{\mathbb{R}^{N-1}} w_n(z) dz_1 \cdots dz_{s-1} dz_{s+1} \cdots dz_N = 1,$$

have the property

$$\lim_{n \to \infty} w_n(z) = \sqrt{\frac{1}{(2\pi)^{N-1} p_1 \cdots p_N}}\, e^{-Q(z)}.$$

Moreover, with the notation

$$W_n(z) := \int_{-\infty}^{z_1} \cdots \int_{-\infty}^{z_{s-1}} \int_{-\infty}^{z_{s+1}} \cdots \int_{-\infty}^{z_N} w_n(y)\, dy_1 \cdots dy_{s-1} dy_{s+1} \cdots dy_N,$$

we get

$$\lim_{n \to \infty} W_n(z) = \sqrt{\frac{1}{(2\pi)^{N-1} p_1 \ldots p_N}} \int_{-\infty}^{z_1} \cdots \int_{-\infty}^{z_{s-1}} \int_{-\infty}^{z_{s+1}} \cdots$$
$$\cdots \int_{-\infty}^{z_N} e^{-Q(y)}\, dy_1 \cdots dy_{s-1} dy_{s+1} \cdots dy_N. \quad (31)$$

Von Mises deduced this theorem (which was in principle known since Laplace (1774, see [2, pp. 167–180])) from a more general theorem on the asymptotic behavior of products of densities. Also in modern expositions (see, for example, Lehmann & Casella [3, sec. 6.5]), the theorem is treated as a special case of a more general situation.

Appendix B

Lemma 6 *If $z_0 \leq \ldots \leq z_n$ and $\alpha_0, \ldots, \alpha_n$ are non-negative reals numbers such that $z_0 = \sum_{i=0}^{n} \alpha_i$, and if f is a concave function defined on a sufficiently large interval, then*

$$\sum_{i=1}^{n} (f(z_i + \alpha_i) - f(z_i)) \leq f(z_0) - f(\alpha_0).$$

Proof: Concave functions can be characterized by the following *decreasing differences property*: if $x \leq z$ and $\alpha \geq 0$, then

$$f(x + \alpha) - f(x) \geq f(z + \alpha) - f(z).$$

Applying this, for each $i \in \{1, \ldots, n\}$, to $x_i := \sum_{k=0}^{i-1} \alpha_k$ and z_i and α_i, and summing the inequalities, one gets

$$f(z_0) - f(\alpha_0) = \sum_{i=1}^{n} (f(x_i + \alpha_i) - f(x_i)) \geq \sum_{i=1}^{n} (f(z_k + \alpha_i) - f(z_i)). \quad \blacksquare$$

Lemma 7 *Let $0 \leq z_1 \leq \ldots \leq z_n$ be real numbers, and let f be non-negative concave function defined on $[0, z_n]$ satisfying $f(0) = 0$. Then, for any integer k, we have the implication*

$$k \cdot z_n \leq \sum_{i=1}^{n} z_i \quad \Longrightarrow \quad k \cdot f(z_n) \leq \sum_{i=1}^{n} f(z_i).$$

Proof: By induction on n, the case $n = 1$ being an immediate consequence of the non-negativity of f.

For the induction step $n \mapsto n+1$, let $0 \leq z_0 \leq \ldots \leq z_n$ and an integer k be given such that

$$k \cdot z_n \leq \sum_{i=0}^{n} z_i. \tag{32}$$

Then we have to show that

$$k \cdot f(z_n) \leq \sum_{i=0}^{n} f(z_i). \tag{33}$$

To this aim we consider non-negative real numbers $\alpha_0, \ldots, \alpha_{n-1}$, and, in order to simplify the notation, $a_n = 0$. For $y_i := z_i + \alpha_i$ $(1 \leq i \leq n)$ we demand

$$z_0 = \sum_{i=0}^{n} \alpha_i, \qquad y_1 \leq \ldots \leq y_n = z_n. \tag{34}$$

Two cases are important. Case 1: $z_0 \leq \sum_{i=1}^{n-1}(z_n - z_i)$. In this case we can choose α_i in accord with (34) such that $\alpha_0 = 0$.
Case 2: $z_0 > \sum_{i=1}^{n-1}(z_n - z_i)$. In this case we choose $\alpha_0 > 0$ and $\alpha_i = z_n - z_i$ for $1 \leq i \leq n-1$.
In the first case we have:

$$\begin{aligned}
k \cdot f(z_n) = k \cdot f(y_n) &\leq \sum_{i=1}^{n} f(y_i) && \text{(induction hypothesis))} \\
&= \sum_{i=1}^{n}(f(z_i + \alpha_i) - f(z_i)) + \sum_{i=1}^{n} f(z_i) \\
&\leq f(z_0) - f(\alpha_0) + \sum_{i=1}^{n} f(z_i) && \text{(by lemma 6)} \\
&= \sum_{i=0}^{n} f(z_i) && \text{(because } f(0) = 0\text{)}.
\end{aligned}$$

For dealing with the second case, first note that the sum (32) is bounded above by $(n+1)z_n$, whence $k \leq n+1$. The case $k = n+1$ is only possible if $z_0 = \ldots = z_n$, in which case (33) is also an equality. Hence, we can restrict our attention to $k \leq n$ (note that here we make use of the assumption that k is an integer). Now the computation runs as follows:

$$\begin{aligned}
k \cdot f(z_n) \leq n \cdot f(z_n) &= \sum_{i=1}^{n} f(y_i) && (y_1 = y_2 = \cdots = y_n = z_n) \\
&\leq f(\alpha_0) + \sum_{i=1}^{n} f(y_i) && \text{(because } f \geq 0\text{)} \\
&= f(\alpha_0) + \sum_{i=1}^{n}(f(z_i + \alpha_i) - f(z_i)) + \sum_{i=1}^{n} f(z_i) \\
&\leq f(\alpha_0) + (f(z_0) - f(\alpha_0)) + \sum_{i=1}^{n} f(z_i) && \text{(by lemma 6)} \\
&= \sum_{i=0}^{n} f(z_i),
\end{aligned}$$

which is the desired inequality (33). ∎

References

[1] S. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, pages 310–318, San Francisco, 1996. Morgan Kaufmann Publishers.

[2] Anders Hald. *A History of Mathematical Statistics from 1750 to 1930*. Wiley, New York, 1998.

[3] E. L. Lehmann and George Casella. *Theory of Point Estimation*. Springer, New York, second edition, 1998.

[4] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[5] E.S. Ristad and P.N. Yianilos. Learning string edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(2):522–532, 1998.

[6] Richard von Mises. Fundamentalsätze der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 4:1–97, 1919.

[7] Günther Wirsching. Speech recognition by statistical language model using square-root smoothing. European Patent 1887562, 2008.

[8] Günther Wirsching. Channel modeling and Levenshtein distances with context-dependent weights. Katholische Universität Eichstätt-Ingolstadt, Preprint Mathematik, 2010. `http://edoc.ku-eichstaett.de/5598/1/cmldcw.pdf`.

Email:
`guenther.wirsching@ku-eichstaett.de`
`hans.fischer@ku-eichstaett.de`