The Score-Distributional Threshold Optimization for Adaptive Binary Classification Tasks

Avi Arampatzis
Computing Science Institute
University of Nijmegen, Postbus 9010
6500 GL Nijmegen, The Netherlands
avgerino@cs.kun.nl

André van Hameren
Institute for Theoretical Physics
University of Nijmegen, Toernooiveld 1
6525 ED Nijmegen, The Netherlands
andrevh@sci.kun.nl

ABSTRACT

The thresholding of document scores has proved critical for the effectiveness of classification tasks. We review the most important approaches to thresholding, and introduce the *score-distributional (S-D) threshold optimization* method. The method is based on score distributions and is capable of optimizing any effectiveness measure defined in terms of the traditional contingency table.

As a byproduct, we provide a model for *score distributions*, and demonstrate its high accuracy in describing empirical data. The estimation method can be performed incrementally, a highly desirable feature for adaptive environments. Our work in modeling score distributions is useful beyond threshold optimization problems. It directly applies to other retrieval environments that make use of score distributions, e.g., distributed retrieval, or topic detection and tracking.

The most accurate version of S-D thresholding — although incremental — can be computationally heavy. Therefore, we also investigate more practical solutions. We suggest practical approximations and discuss adaptivity, threshold initialization, and incrementality issues. The practical version of S-D thresholding has been tested in the context of the TREC-9 Filtering Track and found to be very effective [2].

1. INTRODUCTION

Traditional retrieval systems display documents in a decreasing order of their *scores* with respect to a request. A score may correspond to the probability of relevance of the document, or to some other similarity measure. The user is supposed to go down such a ranked list of documents, and stop at some point determined by the satisfaction (or dissatisfaction) of her request. In some retrieval applications, however, rankings are not enough.

In *binary classification* tasks, e.g. document filtering, a decision should be made for every document whether it belongs to a class or not. If a system is supposed to operate over long periods of time, the interaction between the system and users should be minimized due to cost factors. Decisions such as where to "cut" a ranked list have to be made automatically by the system. In some cases, decisions are required to be taken as soon as a document arrives, therefore ranked lists are not even possible. These considerations suggest the *thresholding* of document scores.

The degree of satisfaction or dissatisfaction of a user may be expressed by an *effectiveness measure*, and the goal of a system is to *optimize* this measure. Thresholding strongly affects effectiveness, and there is no single threshold which optimizes all effectiveness measures. As an example consider two users: the first user values every relevant document as 1 unit of currency, the second user as 10 units, while a non-relevant document costs to both users 1 unit. Assuming that a ranked list has more and more non-relevant documents at lower ranks, the gain of the first user will peak at a higher rank than that of the second. Thus, the corresponding *optimal thresholds* are different.

A classification system operating over long periods of time may accumulate history, e.g. documents and maybe relevance judgments. History can be used to alter the classification model, in order to make better predictions in the future. Systems that alter the classification model as a response to the history are called *adaptive*. Adaptive systems should be able to perform updates in a limited number of calculations and memory. These practical considerations suggest that only a portion of the history should be retained, and algorithms ought to be implemented *incrementally*.

2. OPTIMIZING THRESHOLDS

Let us assume that a set of n documents has been judged by a user, and that r of them have been found relevant to a certain request. Then, the same set of documents is given to a classification system which makes a decision for each document whether to retrieve it or not. All possible four combinations of the user's judgments and the system's decisions can be summarized (quite traditionally) in the contingency Table 1.

system's decision	**user's judgment** relevant	non-relevant
retrieved	R_+	N_+
non-retrieved	R_-	N_-
total	r	$n-r$

Table 1: The traditional contingency table.

SIGIR'01 September 9–12, 2001, New Orleans, Louisiana, USA.

The variables R_+, N_+, R_-, N_-, refer to the number of documents in each category. Effectiveness measures in retrieval tasks are usually defined as functions of these variables. Through the years, a wide range of effectiveness measures have been defined, e.g., precision, recall, the F measure, error rate, and utility, just to name a few popular ones.

2.1 The Probability Thresholding Principle

From the point of view of optimizing measures, D. Lewis in [9] has formulated the *probability thresholding principle* (PTP):

> *"For a given effectiveness measure, there exists a threshold p, $0 \leq p \leq 1$, such that for any set of items, if all and only those items with probability of class membership greater than p are assigned to the class, the expected effectiveness of the classification will be the best possible for that set of items."*

The PTP is a strengthening of the *probability ranking principle* [10] to address the limitations of the latter in classification environments.

The PTP creates two categories of effectiveness measures: measures for which the PTP applies, and measures for which it does not. For the former measures, optimizing a threshold is *theoretically* trivial (we will see the practical difficulties soon). A threshold on probability of relevance can be set once, and the system is guaranteed to exhibit optimal effectiveness in the future, no matter what the distribution of probabilities of relevance for documents is.

As an example, let us consider the family $U_{(\lambda_1,\lambda_2,\lambda_3,\lambda_4)}$ of *linear utility functions*:

$$U_{(\lambda_1,\lambda_2,\lambda_3,\lambda_4)} = \lambda_1 R_+ + \lambda_2 N_+ + \lambda_3 R_- + \lambda_4 N_- \ , \qquad (1)$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ denote the gain or cost associated with each document that falls under the corresponding category. The optimal probability threshold associated with any of the above functions has been shown in [7] to be:

$$p = \frac{\lambda_2 - \lambda_4}{(\lambda_3 - \lambda_1) + (\lambda_2 - \lambda_4)} = \frac{1}{1+\lambda} \ , \qquad (2)$$

where

$$\lambda = \frac{\lambda_3 - \lambda_1}{\lambda_2 - \lambda_4} \ . \qquad (3)$$

Since the optimal threshold depends only on the measure, the PTP holds.

Practically, such probabilistic thresholds are difficult to apply. The main reason is that even probabilistic retrieval models do not obtain the actual probabilities of relevance for documents. Traditional probabilistic models make extensive use of *order-preserving* transformations (some of which are difficult to reverse) of probabilities of relevance. Any such transformation does not affect ranked retrieval, but makes formulae like (2) practically useless, unless a way is found to reverse the transformations. A transformation reversal strategy has been adopted by the Okapi probabilistic system with rather successful results [11].

For non-probabilistic retrieval models, however, how to turn a similarity score into a probability of relevance is still a fair question. In any case, optimizing a measure for which the PTP does not hold (e.g. for the F measure [9]) requires other considerations. A method based on *score distributions*, irrespective of what a score is, would be more general and valid for any measure or retrieval model.

2.2 The Straightforward Method

There exists a distributional procedure, which we will call the *straightforward empirical* method, that is capable of finding an optimal threshold on training data. It consists of the following steps:

- calculate the scores of all training documents,
- rank them,
- calculate the effectiveness measure at every position of the rank,
- go down the rank and find the position where the effectiveness measure is optimal,
- set the threshold somewhere between the score that corresponds to the position above and the next one.

The technique implicitly considers the density of relevant to the non-relevant documents and the spread of their scores. It has been applied many times before and, given sufficient training data, works well (see e.g. [12]).

Although the straightforward empirical method seems like a perfect choice for optimizing thresholds in classification tasks, its drawbacks become apparent when adaptivity is required. Firstly, there is no known way to implement it incrementally. The scores of all accumulated training documents have to be re-calculated after every query update, therefore document buffers are required. The fixed memory model requirement of practical systems means that buffers should be of limited size, thus some documents have to be discarded as the history grows. This may have a negative impact on the estimation accuracy, especially when the convergence of classifiers is more important than responsiveness[1]. Secondly, the method gives absolutely no prediction of the optimal threshold when there is no relevance information, and it is bound to be very inaccurate with sparse relevance data.

Our S-D method has the following advantages over the above empirical technique.

1. It allows for better incrementality, retaining accuracy. Most of the quantities it needs for the estimation can be updated incrementally when new data become available.
2. It can give better predictions of where the optimal threshold may be, when there is sparse or even no relevance information.
3. It uses the statistical properties of the scores rather than the actual values. Therefore, the estimation of the optimal threshold may generalize better to unseen documents.

[1]Responsiveness of classifiers is required when relevance drifts exist. In such cases, old training data may be discarded more safely, since their relevance judgment was valid at the time it was generated and may not correspond to now. Such considerations and others can be found in [3, 1].

3. THE S-D THRESHOLD OPTIMIZATION

The S-D threshold optimization method can be applied for any effectiveness measure of the form $M(R_+, N_+, R_-, N_-)$, i.e. M is any function of the variables of the contingency Table 1. The optimization is based on the score distributions of relevant and non-relevant documents, and on their relative density in a document set.

Let us assume that the scores of relevant documents are distributed with a probability density function $P_{\rm r}(x)$. Then, the quantity $rP_{\rm r}(x)\,dx$ gives the number of relevant documents with scores in the range $[x, x+dx)$. The number of relevant documents which score above a threshold θ is

$$R_+(\theta) = r\int_\theta^{+\infty} P_{\rm r}(x)\,dx\ . \tag{4}$$

The number of non-relevant documents with scores above θ is similarly defined as

$$N_+(\theta) = (n-r)\int_\theta^{+\infty} P_{\rm nr}(x)\,dx\ , \tag{5}$$

where $P_{\rm nr}(x)$ the probability density function of the score distribution of non-relevant documents. The numbers of relevant non-retrieved and non-relevant non-retrieved documents for θ are given respectively by

$$R_-(\theta) = r - R_+(\theta)\ , \tag{6}$$
$$N_-(\theta) = (n-r) - N_+(\theta)\ . \tag{7}$$

Using the last four equations, M can be written as a function of θ: $M\left(R_+(\theta), N_+(\theta), R_-(\theta), N_-(\theta)\right)$.

Optimizing M means either maximizing or minimizing it (depending on whether larger M means better effectiveness or the other way around), therefore the optimal threshold is a solution of

$$\frac{dM\left(R_+(\theta), N_+(\theta), R_-(\theta), N_-(\theta)\right)}{d\theta} = 0\ . \tag{8}$$

In order to solve this equation for a given M, we first need to define the probability densities $P_{\rm r}(x)$ and $P_{\rm nr}(x)$. We will model these distributions in Sec.4, and suggest practical approximations in Sec.5.1 and 5.3.

In most cases, (8) does not have analytical solutions, so it has to be solved numerically. For linear measures, however, it simplifies greatly since the integrals cancel out with the derivative. For example, for linear utility functions (Eq. (1)), after a few calculations (8) becomes

$$\lambda\rho P_{\rm r}(\theta) = P_{\rm nr}(\theta)\ , \tag{9}$$

where λ is given by (3), and $\rho = \frac{r}{n-r}$ is the *relative density* of relevant to the non-relevant documents.

The probability $P(\text{rel}|s)$ of a document with score s to be relevant may be expressed as

$$P(\text{rel}|s) = \frac{rP_{\rm r}(s)}{rP_{\rm r}(s) + (n-r)P_{\rm nr}(s)}\ . \tag{10}$$

The probability of relevance at $s=\theta$ can be calculated by using (9) on (10). The result is $P(\text{rel}|\theta) = \frac{1}{1+\lambda}$, i.e. the same as (2). Obviously, our method may be used, via (10), to reverse scores into probabilities of relevance, however, we do not see the need to do that since we can calculate the optimal threshold in the first place.

4. SCORE DISTRIBUTIONS

C. Baumgarten in [4] has modeled score distributions, using the mean and deviation of the data, with a *gamma distribution* shifted by the minimum score. The motivation for using a gamma distribution has been empirical, but the approach has worked out well. We will rather set out to build a model from scratch.

Let us represent a query by an m-tuple $\boldsymbol{q} = [q_1, \ldots, q_m]$, where q_i is a value that corresponds to the term i. A document is represented similarly, using the same set of terms, as $\boldsymbol{\omega} = [\omega_1, \ldots, \omega_m]$. The values of the terms in documents depend on a weighting scheme W. Subsequently, $\boldsymbol{q}$ and W together determine the structure of the document space. We will specify W only qualitatively such as: the larger the similarity of a document to the query, the larger the document score defined through the linear function of document weights:

$$\langle \boldsymbol{q}, \boldsymbol{\omega}\rangle = \boldsymbol{q} * \boldsymbol{\omega} = \sum_i q_i\omega_i\ . \tag{11}$$

Represented as m-tuples, documents and query are obviously points in $\mathbb{R}^m$.

Our aim is to calculate the distribution of the scores of a general class C of documents. Since the score of a document is a linear combination of its components, the score distribution can be derived from the distribution of the documents in $\mathbb{R}^m$. This distribution can be represented by a probability measure P_m on $\mathbb{R}^m$. For every *convex subset*[2] $A \subset \mathbb{R}^m$, the number $\mathsf{P}_m(A)$ gives the fraction of documents from C for which their m-tuples are in A. Although a real-life set of documents is countable, we represent it by the *continuous* space $\mathbb{R}^m$. The large number of different documents makes this a reasonable approximation.

Of course, the distribution of documents does not have to be smooth in $\mathbb{R}^m$, and all documents may be restricted to a hyper-surface in $\mathbb{R}^m$ of lower dimension than m, say $m-1$. Strictly speaking, we should then define a measure P_{m-1} on this (curved) lower dimensional space. We, however, prefer to formulate everything in $\mathbb{R}^m$, and to put possible constraints in P_m with the help of *Dirac* δ *distributions*[3].

Let us denote $[\boldsymbol{\alpha}, \boldsymbol{\beta}) = [\alpha_1, \beta_1) \times [\alpha_2, \beta_2) \times \cdots \times [\alpha_m, \beta_m)$ and $\mathsf{P}_m(d\boldsymbol{\omega}) := \mathsf{P}_m(\,[\boldsymbol{\omega}, \boldsymbol{\omega} + d\boldsymbol{\omega})\,)$. Given P_m, the characteristic function ϕ of the score distribution is given by

$$\phi(t) = \mathsf{E}(\,e^{\imath t\langle \boldsymbol{q},\boldsymbol{\omega}\rangle}\,) = \int_{\mathbb{R}^m} e^{\imath t\langle \boldsymbol{q},\boldsymbol{\omega}\rangle}\,\mathsf{P}_m(d\boldsymbol{\omega})\ , \tag{12}$$

and the probability density of the scores of class C is given by the Fourier transform of ϕ [8]:

$$P_C(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-\imath xt}\phi(t)\,dt\ . \tag{13}$$

[2]The convexity of A is a fair requirement. Suppose you do not demand A to be convex, for example take A to be such that it consists of tiny balls around the points of the documents, connected by very narrow tubes. Then, P_m will look like a collection of peaks at the points of the documents. In order to smooth these peaks out and get a nice continuum limit, the convexity of the subspaces A is required.

[3]For example, if all documents happen to be distributed on a hyper-sphere in $\mathbb{R}^m$ with center $[0, 0, \ldots, 0]$ and radius R, then $\mathsf{P}_m(d\boldsymbol{\omega}) = P(\boldsymbol{\omega})\delta(\|\boldsymbol{\omega}\| - R)\,d\boldsymbol{\omega}$, where P is a positive function on $\mathbb{R}^m$ such that $\int_{\mathbb{R}^m} P(\boldsymbol{\omega})\delta(\|\boldsymbol{\omega}\| - R)\,d\boldsymbol{\omega} = 1$. The δ distribution restricts the measure to be non-zero only for documents that have lengths equal to R.

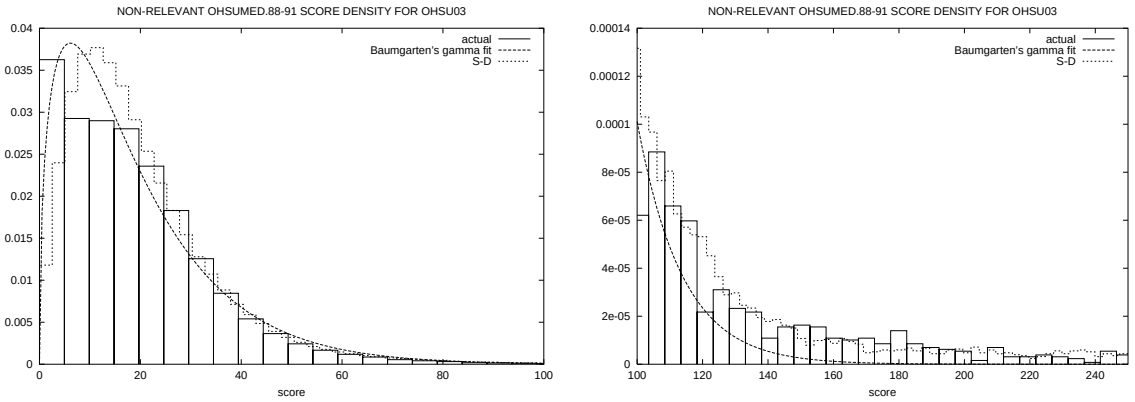

Figure 1: Body (left) and tail (right) of the score density of non-relevant documents. Zero scores are excluded.

In the formulation above, the components ω_i of the documents can be considered random variables, and the score is a linear combination of these random variables

$$S_m = \sum_{i=1}^{m} X_i \ , \quad X_i = q_i \omega_i \ . \tag{14}$$

We will make the (common in IR) assumption that

ASSUMPTION 1. *the components ω_i of documents are distributed independently.*

For the measure P_m, this means that it factorizes over the components of $\mathbb{R}^m$, i.e. there are m 1-dimensional measures p_i so that we can write

$$\mathsf{P}_m(d\boldsymbol{\omega}) = \prod_{i=1}^{m} \mathsf{p}_i(d\omega_i) \ . \tag{15}$$

As a result of this and the linearity of the score as function of document components, the characteristic function can be written as a product of characteristic functions of the components:

$$\phi(t) = \prod_{i=1}^{m} \phi_i(q_i t) \ , \quad \phi_i(q_i t) = \int_{-\infty}^{\infty} e^{\imath q_i t \omega_i} \, \mathsf{p}_i(d\omega_i) \ . \tag{16}$$

In order to construct the 1-dimensional measures p_i, we observe that weighting schemes usually are such, that if a term does not appear at all in a document, then this term gets weight zero. We relate the probability of term i to appear in a document directly to its document frequency across class C by defining

$$\varepsilon_i = \frac{\text{number documents in } C \text{ containing term } i}{\text{total number of documents in } C} \ , \tag{17}$$

and we call it the *term probability* (TP). Consequently, the measure p_i will have the form

$$\mathsf{p}_i(\, \omega_i \leq x \,) = (1 - \varepsilon_i)\vartheta(x) + \varepsilon_i F_i(x) \ , \tag{18}$$

where ϑ is the step function, and F_i is some probability distribution function (PDF) which depends on W. In the simplest case of binary weighted document terms, $F_i(x) = \vartheta(x-1)$, $\forall i$. In general, F_i can be derived directly from the W being used, or estimated empirically from a dataset.

So far, we have built a model for the score distribution of a general class C of documents. The model is capable of calculating the distribution from TPs and $\boldsymbol{q}$. The only assumption we have made is that of independence of term occurrences. We have left open the form of functions F_i; these should be defined according to the W used.

Turning to the independence assumption, our model will more likely work better when there are less violations of the assumption. This suggest a small number of dimensions m, or that the model should be used for document classes consisting of documents that have a small number of components matched with the query, e.g. the class of non-relevant documents $C_{\rm nr}$. Dependencies blow up the scores. Our model, however, allows us to take the dependencies indirectly into account, through the functions F_i; these can be adjusted accordingly to compensate for the score blow-ups, as we will see next.

4.1 Evaluation of the Model

Fig.1 shows the empirical score distribution of non-relevant documents, Baumgarten's gamma distribution fit, and the density calculated by our model. Our S-D curve is calculated with a *Monte Carlo* method [15], which is why it is plotted with steps. The Rocchio-expanded query has around 400 dimensions. Training documents were *Ltu* weighted, while test documents were *Lu* weighted [13].

We approximated *Lu* and the dependencies introduced due to the large number of dimensions by

$$F_i(x) = F(x) = \frac{\log(x) - \log(a)}{\log(b) - \log(a)} \ , \quad 0 < a < b \ , \quad \forall i \ . \tag{19}$$

This means that the density function coming with F behaves as $1/x$ between a and b. We used the values that give a good fit with the empirical data: $a = 0.1$ and $b = 3.5$. We want to stress that, according to our observations, these parameters can be taken constant for different queries of approximately the same length.

Eq.(19) certainly does not correspond to *Lu* weights. It is just an *ad hoc* formula to demonstrate how robust our model is: we have effectively obtained a very good fit on the empirical data, using the same F for all terms, and the effect of dependencies has turned out to be directly related to the *number* of dimensions, no matter which ones. The gamma distribution, nevertheless, gives a surprisingly good fit over a range of queries and dimensionalities. But our model is more accurate exactly where this is needed: *on the tail*.

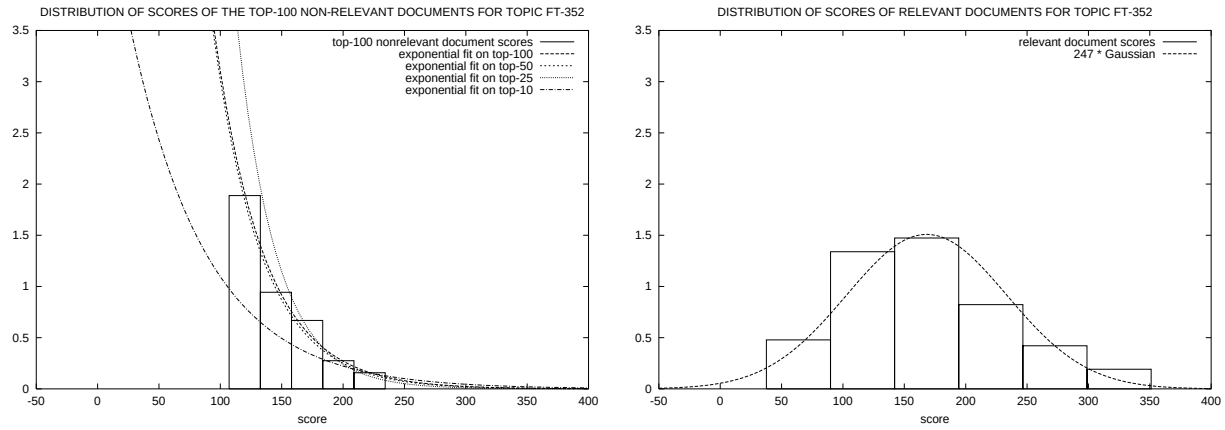

Figure 2: Empirical score distributions and the corresponding exponential (left) and Gaussian (right) fit.

5. PRACTICAL S-D OPTIMIZATION

So far, we have worked out an accurate optimization at all costs. As a result, it may be computationally heavy to calculate the score densities $P_{\rm r}$ and $P_{\rm nr}$ using the model we have described in Sec.4. Moreover, the model of Sec.4 may break down for $P_{\rm r}$ due to the increased number of dependencies. In any case, the goal of threshold optimization is to improve filtering, and too much of a threshold accuracy may not pay off in effectiveness (this still remains to be seen). Let us see how the optimization can be applied more efficiently without sacrificing too much accuracy.

5.1 Gaussian Limits?

In order to simplify the calculation of the score densities, it is sensible to look if a Central Limit Theorem [8] applies to S_m (Eq.(14)) in the limit of a large number of dimensions m, and that the score distribution becomes Gaussian in this limit. If the answer to the question *whether* a Gaussian limit appears is yes, then the next question is *when* it appears, i.e., for which values of m. Which values of m can be considered large?

For $P_{\rm nr}$, we show in App.A that a Gaussian limit is not likely, and if it appears, then only at a very slow rate with m. Empirically, we have never seen Gaussian shapes even for *all* dimensions resulted from massive expansion of queries. In App.B, we prove that a Gaussian limit appears for $P_{\rm r}$. Furthermore, we show that the distribution approaches the Gaussian quickly, such that corrections go to zero as $1/m$. Empirically, Gaussian shapes form at around $m = 250$ (Fig.2, right).

5.2 The Curse of Dimensionality

To ensure a Gaussian central limit for $P_{\rm r}$, high dimensionality is required. Obviously, long queries can only be obtained with massive expansion through e.g. relevance feedback. One could argue against high dimensionality for efficiency reasons or due to the increased term dependencies introduced. Massive query expansion, however, has been shown to be effective [5]. Moreover, long queries are necessary when tracking relevance drifts, which are likely to occur in the retrieval environments we consider [3]. Above all, setting the thresholds right has proved to be critical for effectiveness in classification environments.

We do not recommend giving up on high dimensionality, since shorter queries may give zero scores for relevant documents truncating $P_{\rm r}$ at zero. Not only it is not obvious how to estimate the parameters of a truncated distribution, but also our empirical data seem too irregular to be modeled by anything. A Gaussian limit for $P_{\rm r}$ is convenient, simplifying the calculations greatly.

5.3 Approximation with an Exponential

We have proved that $P_{\rm nr}$ does not have a Gaussian limit —at least not in a usable number of dimensions— so the Fourier transform method (Eq.(13)) seems unavoidable. However, a great (but *ad hoc*) simplification would be to fit a simple exponential of the form $c_1 \exp(-c_2 x)$ on the empirical non-relevant score distribution. This approach has worked out well in [2], using a buffer of the top-50 scoring non-relevant documents and 5 bins (Fig.2, left).

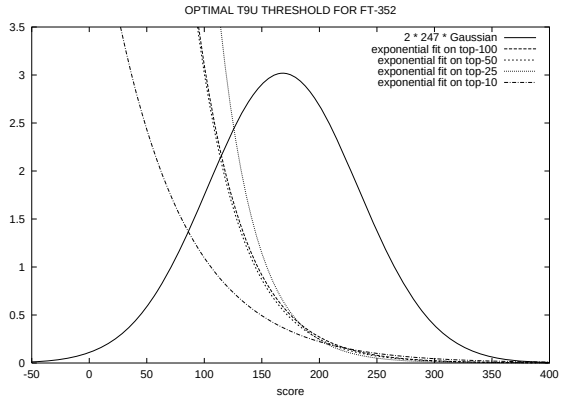

Figure 3: The optimal T9U threshold.

An extra bonus of using an exponential $P_{\rm nr}$ is that, for linear measures, (9) can be solved analytically (using a Gaussian $P_{\rm r}$). Fig.3 shows the optimal T9U[4] threshold, which is just the score at which the densities $P_{\rm r}$ and $P_{\rm nr}$, weighed as λr and $n - r$ respectively (Eq.(9)), intersect each other. The complete analytical solution for optimizing linear utility functions can be found in [2] or [1, pg. 53–54]. For nonlinear measures, however, (8) still has to be solved numerically.

[4]T9U is a linear utility with $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (2, -1, 0, 0)$.

5.4 Threshold Initialization

To estimate a Gaussian $P_{\rm r}$, our method so far relies on relevance information. ($P_{\rm nr}$ can be constructed with no relevance information, using (13) and TPs calculated on *all* documents C_n. Those TPs are a good approximation of the TPs of the class of non-relevant documents $C_{\rm nr}$, since $n >> r$.) How should $P_{\rm r}$ be initialized when there is no relevance information?

The query itself can give an estimate of where $P_{\rm r}$ lies, e.g. $||\boldsymbol{q}||^2$ can be seen as the maximum relevant score. Some reasonable assumption (by taking $P_{\rm nr}$ into account as well) for the standard deviation $\sigma_{\rm r}$ of $P_{\rm r}$ can produce a usable curve, e.g. through an equation $\mu_{\rm r} = ||\boldsymbol{q}||^2 - 3\sigma_{\rm r}$, where $\mu_{\rm r}$ is the mean of $P_{\rm r}$.

5.5 Incrementality

Let us see how the Gaussian–exponential model can be applied incrementally in adaptive environments. A Gaussian is defined by its mean and deviation. In general, means and deviations can be updated incrementally. In our context, however, every query update causes the scores of previously seen documents to change, suggesting that all scores should be recalculated. Assuming a *static* W, in the sense that document weights do not depend on any statistics external to documents (e.g. documents are only *tf*-weighted), it has been shown in [6, 2] that

$$\mu_{\rm r} = \frac{1}{r}\sum_{i=1}^{r}\langle \boldsymbol{q}, \boldsymbol{\omega}_i\rangle = \frac{1}{r}\langle \boldsymbol{q}, \sum_{i=1}^{r}\boldsymbol{\omega}_i\rangle \ . \qquad (20)$$

Obviously, the sum of relevant document tuples is sufficient and can be updated incrementally.

The variance $\sigma_{\rm r}^2$ can be calculated via $\sigma_{\rm r}^2 = \mu_{\rm r}^{(2)} - \mu_{\rm r}^2$, where the mean of the squared scores is given by

$$\mu_{\rm r}^{(2)} = \frac{1}{r}\sum_{i=1}^{r}\langle \boldsymbol{q}, \boldsymbol{\omega}_i\rangle^2 = \frac{1}{r}\sum_{jk} q_j \left(\sum_{i=1}^{r}\omega_{ij}\omega_{ik}\right) q_k \ , \qquad (21)$$

where e.g. ω_{ij} is the value of the jth component of the ith document. The proof of (21) is given in [1, pg. 144]. The sum in the parenthesis can be represented by a 2-dimensional matrix $\mathbf{o}$ with components

$$o_{jk}^{(r)} = \sum_{i=1}^{r}\omega_{ij}\omega_{ik} \ , \qquad (22)$$

and it can be updated as $o_{jk}^{(r+1)} = o_{jk}^{(r)} + \omega_{(r+1)j}\omega_{(r+1)k}$, upon the arrival of document $\boldsymbol{\omega}_{r+1}$.

The exponential fit for $P_{\rm nr}$ requires a small document buffer to hold the top-scoring non-relevant (retrieved) documents, because all scores should be recalculated. If the buffer is full when a new non-relevant document is retrieved, the approach of ranking the buffered documents and discarding the lowest-scoring one has worked out well in [1, 2].

5.6 Adaptivity and Soft Thresholds

A special problem that shows up in adaptive environments is that relevance information is becoming available only for documents retrieved. This may invalidate the score statistics required, and lead a system to a *selectivity trap* [11, 3]. For instance, estimating a Gaussian from data which do not include its left tail (these are the data below the threshold), may overestimate the threshold, retrieving no more documents.

A solution would be to use a *soft probabilistic threshold*, i.e. a document that scores at s, $s < \theta$, may still be retrieved by sampling it with a probability $P(\text{rel}|s)$ given by (10). Of course, the statistic that a document retrieved like this provides, should be weighted as $1/P(\text{rel}|s)$. In this way, score statistics can be maintained more accurately, and selectivity traps can be avoided. The idea remains to be tested.

6. CONCLUSIONS

We have developed a novel method for optimizing thresholds, namely, the *score distributional (S-D) threshold optimization*. The method is capable of optimizing any effectiveness measure defined in terms of the contingency Table 1. The analysis we have provided is general enough to apply to a range of retrieval models, from probabilistic to vector space. Moreover, S-D thresholding can be applied incrementally, a highly desirable feature for adaptive environments.

S-D thresholding is based on score distributions, therefore we have developed models for their estimation. We have provided a range of choices, from very accurate and computationally expensive to practical and less expensive approximations. Whether our most accurate model for scores (Sec.4) pays off in classification effectiveness by providing better thresholding still remains to be seen.

In our attempts to approximate inexpensively score distributions, we have proved that the distribution of relevant document scores has a Gaussian limit that shows up in a practically usable number of query dimensions (around 250). We have also proved that the distribution of non-relevant document scores does not have a Gaussian limit (at least not in a usable number of dimensions), however, we have empirically found that its right tail can be very well approximated with an exponential. This practical Gaussian–exponential version of the S-D threshold optimization has been tested in the context of the TREC-9 Filtering Track and presented exceptional end-results [2].

At any rate, our work in modelling scores is also useful beyond threshold optimization. It is directly applicable to other retrieval environments that make use of such distributions, e.g., distributed retrieval [4], or topic detection and tracking [14]. For instance, it can be employed (via Eq.10) to reverse document scores into probabilities of relevance, giving a way of combining the output of several search engines into a single ranking.

7. REFERENCES

[1] A. Arampatzis. *Adaptive and Temporally-dependent Document Filtering*. PhD thesis, University of Nijmegen, Nijmegen, The Netherlands, 2001. Available from `www.cs.kun.nl/~avgerino`.

[2] A. Arampatzis, J. Beney, C. Koster, and T. van der Weide. Incrementality, Half-Life, and Threshold Optimization, for Adaptive Document Filtering. In *The Nineth Text REtrieval Conference (TREC-9)*, Gaithersburg, MD, November 13–16 2000. NIST.

[3] A. Arampatzis and T. van der Weide. Document Filtering as an Adaptive and Temporally-dependent Process. In *Proceedings of the BCS-IRSG European Colloquium on IR Research*, Darmstadt, Germany, April 4–6 2001.

[4] C. Baumgarten. A Probabilistic Solution to the Fusion Problem in Distributed Information Retrieval. In

Proceedings of the 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, 1999.

[5] C. Buckley, G. Salton, and J. Allan. The Effect of Adding Relevance Information in a Relevance Feedback Environment. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland, June 1994.

[6] J. Callan. Learning While Filtering Documents. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, August 1998.

[7] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. Wiley-Interscience, New York, 1973.

[8] R. Laha and V. Rohatgi. *Probability Theory*. John Wiley & Sons, New York, 1979.

[9] D. Lewis. Evaluating and optimizing autonomous text classification systems. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, July 1995.

[10] S. Robertson. The Probability Ranking Principle. *Journal of Documentation*, 33(4):294–304, December 1977.

[11] S. Robertson and S. Walker. Threshold Setting in Adaptive Filtering. *Journal of Documentation*, 56:312–331, 2000.

[12] R. Schapire, Y. Singer, and A. Singhal. Boosting and Rocchio Applied to Text Filtering. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, August 1998.

[13] A. Singhal. AT&T at TREC-6. In *The Sixth Text REtrieval Conference (TREC-6)*, Gaithersburg, MD, November 19–21 1997. NIST.

[14] M. Spitters and W. Kraaij. A Language Modeling Approach to Tracking News Events. In *Notebook of the Topic Detection and Tracking 2000 Workshop*, Gaithersburg, MD, November 2000. NIST.

[15] A. van Hameren. *Loaded Dice in Monte Carlo*. PhD thesis, University of Nijmegen, 2001. Available from `http://arXiv.org/abs/hep-ph/0101094`.

APPENDIX

A. NON-GAUSSIAN LIMIT FOR NON-REL.

Let $C_{\rm nr}$ be the class of non-relevant documents. In order to investigate the behavior of the score distribution for $C_{\rm nr}$, we investigate the *cumulants* [8]. These are defined through the moment generating function $\phi(-\imath t)$ of the score distribution, where ϕ is the characteristic function.

$$K_m^{(r)} := \frac{d^r \log \phi(-\imath t)}{dt^r}\Big|_{t=0} \quad . \tag{23}$$

The first cumulant $K_m^{(1)}$ is equal to the mean of the distribution, and the second cumulant $K_m^{(2)}$ is equal to the variance. For a given random variable S_m with given cumulants $K_m^{(r)}$, the cumulants of the variable

$$\hat{S}_m := \frac{S_m - K_m^{(1)}}{\left(K_m^{(2)}\right)^{\frac{1}{2}}} \quad , \tag{24}$$

shifted such that it has zero mean and unit variance, are given by

$$\hat{K}_m^{(1)} := 0 \quad , \qquad \hat{K}_m^{(r)} := \frac{K_m^{(r)}}{\left(K_m^{(2)}\right)^{\frac{r}{2}}} \quad r \geq 2 \ . \tag{25}$$

For a Gaussian distribution, the logarithm of the moment generating function is given by

$$\log \phi(-\imath t) := t K_m^{(1)} + \frac{t^2}{2} K_m^{(2)} \quad . \tag{26}$$

The above trivially leads to the conclusion that

THEOREM 1. *whenever* $\lim_m \hat{K}_m^{(r)} \to 0$ *for all* $r \geq 3$*, then* $\hat{S} = \lim_m \hat{S}_m$ *is a normal variable, that is, it has a Gaussian probability distribution with zero mean and unit variance.*

This theorem is not the most efficient to prove a Gaussian limit, because it asks for the limiting behavior of *all* cumulants, but it gives a view on how *fast* the limit appears: if the cumulants $\hat{K}_m^{(r)}$ go to zero for large m at a very slow rate, then the probability distribution will start to look Gaussian only for very large m.

Because of the independence assumption, the characteristic function factorizes over the components (Eq.(16)), so that its logarithm becomes a sum over the components of logarithms

$$K_m^{(r)} = \sum_{i=1}^{m} q_i^r \kappa_i^{(r)} \quad , \quad \kappa_i^{(r)} := \frac{d^r \log \phi_i(-\imath t)}{dt^r}\Big|_{t=0} \quad . \tag{27}$$

The moments of the components depend linearly on the TPs: according to (18) we have

$$\mathsf{E}(\omega_i^r) = \int_{-\infty}^{\infty} x^r \, \mathsf{p}_i(dx) = \varepsilon_i \int_{-\infty}^{\infty} x^r \, dF_i(x) \quad . \tag{28}$$

The cumulants can be written as finite sums of products of the moments, so that in this case $\kappa_i^{(r)}$ is a polynomial in ε_i, i.e.

$$\kappa_i^{(r)} = \varepsilon_i F_i^{(r)} + \mathcal{P}_{2,r}(\varepsilon_i) \quad , \quad F_i^{(r)} := \int_{-\infty}^{\infty} x^r \, dF_i(x) \quad , \tag{29}$$

where $\mathcal{P}_{2,r}(\varepsilon)$ denotes a polynomial in ε containing orders 2 to r. The interpretation of the cumulants as an expansion in the TPs makes sense, because $\varepsilon_i \leq 1$ by definition.

Now, we shall try to derive from the constructed model whether the score distribution converges to a Gaussian for large m, and if it does, what the rate of convergence is. In order to achieve this, we want to replace the sum in (27) by m times the average query component times the average cumulant. To do this, we need some more assumptions.

The first one is based on the empirical observation that, whereas the TPs ε_i and the moments $\mathsf{E}(\omega_i)$ and $\mathsf{E}(\omega_i^2)$ of the components vary several orders of magnitude, the ratios $\mathsf{E}(\omega_i)/\varepsilon_i$ and $\mathsf{E}(\omega_i^2)/\varepsilon_i$ vary within only one order of magnitude. Together with (28), the mentioned observation leads to the conclusion that the PDFs F_i do not vary much for the different components, or at least that the variations do not matter much. The important differences between the distributions of the components seems to come from the TPs. We implement this in our model by the assumption that

ASSUMPTION 2. *the PDFs F_i are the same for all components, and equal to a single PDF F.*

In order to determine the rate of convergence, we intent to use Theorem 1, so that we need to determine the behavior of the cumulants for large m. According to (27) and (29), we then need the distributions of the query components (QCs) and the TPs. For both cases, we specify the distribution of the variable by applying a *generalization of Zipf's law*. For QCs let denote

$$\bar{q}_m = \text{the value of the maximal QC.}$$

For every m, there is a mapping $\mathcal{Q}_m$ with $\mathcal{Q}_m(1) = 1$, such that the ordered labeling of the variables satisfies

$$q_i = \bar{q}_m \mathcal{Q}_m(i) \quad \text{for every} \quad i = 1, \ldots, m \ . \tag{31}$$

Zipf's classical law is obtained with $\mathcal{Q}_m(i) = 1/i$. The distribution of the variable has moments

$$q_m^{(r)} = \frac{\bar{q}_m^r}{m} \mathcal{Q}_m^{(r)} \quad , \quad \mathcal{Q}_m^{(r)} := \sum_{i=1}^{m} \mathcal{Q}_m(i)^r \ .$$

By definition, the mapping $\mathcal{Q}_m$ is decreasing with $\mathcal{Q}_m(1) = 1$, so that $\mathcal{Q}_m(i)^{r_1} \geq \mathcal{Q}_m(i)^{r_2}$ for all $i = 1, \ldots, m$ if $r_1 < r_2$ and

$$\mathcal{Q}_m^{(r_1)} \geq \mathcal{Q}_m^{(r_2)} \quad \text{for} \quad r_1 < r_2 \ . \tag{33}$$

Furthermore, all moments exist, also in the limit of $m \to \infty$, since in the worst case we would have $\mathcal{Q}_m(i) = 1$ for all $i = 1, \ldots, m$, so that $q_m^{(r)} = \bar{q}_m^r$. Therefore, we conclude that

$$\mathcal{Q}_m^{(r)} = \mathcal{O}(m) \quad \text{for all} \quad r > 0 \ ,$$

where the $\mathcal{O}$-symbol refers to the behavior with m: we say $a_m = \mathcal{O}(b_m)$ if there is a sequence of numbers c_m such that $|a_m/b_m| < c_m$ for all m, and $\lim_{m\to\infty} c_m$ exists. The sums $\mathcal{Q}_m^{(r)}$ do not have to exist in the limit $m \to \infty$: for example in the classical Zipf case, we have $\mathcal{Q}_m^{(1)} = \log m + \mathcal{O}(1)$.

Exactly the same can be done for the TPs, leading to a maximal value $\bar{\varepsilon}_m$, a decreasing mapping $\mathcal{E}_m$ with $\mathcal{E}_m(1) = 1$ and such that $\varepsilon_i = \bar{\varepsilon}_m \mathcal{E}_m(i)$. The moments of the TPs are denoted $\varepsilon_m^{(r)}$.

At this point, we want to notice that the ordering (33) of the coefficients $\mathcal{E}_m^{(r)}$ corresponds with the ordering of powers of $\bar{\varepsilon}_m$, which supports the approximation to

APPROXIMATION 1. *keep only the lowest order in ε_i for every $i = 1, \ldots, m$ in (29),*

since ε_i is smaller than 1.

The following assumption is based on the empirical observation that QCs and TPs seem to take their values independently: if we order the QCs, and make a plot of the values of the corresponding TPs in this ordering, they seem to jump around randomly. This suggests to assume that

ASSUMPTION 3. *the TPs and the QCs take their values independently of each other,*

so that the average over the TPs can be taken independently of the average of the QCs. Assumption 3 together with (27) and Approximation 1 lead to

$$K_m^{(r)} \overset{m\to\infty}{\longrightarrow} m\, q_m^{(r)} \varepsilon_m^{(1)} F^{(r)} = \frac{1}{m} \bar{q}_m^r \mathcal{Q}_m^{(r)} \bar{\varepsilon}_m \mathcal{E}_m^{(1)} F^{(r)} \ . \tag{35}$$

The interesting ratio of the cumulants is then given by

$$\frac{K_m^{(r)}}{\left(K_m^{(2)}\right)^{\frac{r}{2}}} = \frac{\mathcal{Q}_m^{(r)}}{\left(\mathcal{Q}_m^{(2)}\right)^{\frac{r}{2}}} \times \left(\frac{m}{\bar{\varepsilon}_m \mathcal{E}_m^{(1)}}\right)^{\frac{r}{2}-1} \times \frac{F^{(r)}}{\left(F^{(2)}\right)^{\frac{r}{2}}} \ . \tag{36}$$

According to Theorem 1, the score distribution becomes Gaussian for large m if this final expression vanishes for all $r \geq 3$.

The main use of Assumption 3 is that it enables us to give an estimate of the ratios on the l.h.s. of (36), in which the contribution of the query completely factorizes from the contribution from the document distribution. A possible difference in the rate of convergence between two document classes only appears in the second and the third factor of the r.h.s. of (36).

The contribution from the first factor on the r.h.s. of (36) is determined by the distribution of the QCs. Using (33) and the fact that $\mathcal{Q}_m^{(r)} \geq 1$ for every $r \geq 3$, we see that

$$\lim_{m\to\infty} \frac{\mathcal{Q}_m^{(r)}}{\left(\mathcal{Q}_m^{(2)}\right)^{\frac{r}{2}}} = 0 \quad \Longleftrightarrow \quad \lim_{m\to\infty} \frac{1}{\mathcal{Q}_m^{(2)}} = 0 \ . \tag{37}$$

So the contribution from the QCs only helps towards a Gaussian limit if $\lim_{m\to\infty} \mathcal{Q}_m^{(2)} = \infty$. We observe a behavior of the distribution of the QCs such that $\mathcal{Q}_m(i) \sim (i)^{-\nu}$ with $0.5 < \nu < 1$. For this behavior, only the case of $\nu = 0.5$ would, strictly speaking, lead to the first factor on the r.h.s. of (36) to become zero, as $(\log m)^{r/2}$. We conclude that, if the distribution of the QCs helps towards a Gaussian limit, then only very slowly.

The third factor on the r.h.s. of (36) does not vary with m (by Assumption 2), so that we further only need to look at the second factor, which is determined by the distribution of the TPs. Since $\bar{\varepsilon}_m \leq 1$, only the behavior of the mapping $\mathcal{E}_m$ can help towards a Gaussian limit, and then only if $\lim_{m\to\infty} \mathcal{E}_m^{(1)}$ does not exist. However, we know that $\mathcal{E}_m^{(1)} = \mathcal{O}(m)$, so that the second factor on the r.h.s. of (36) will never go to zero.

We conclude that it is not likely for C_{nr} to show a Gaussian limit, and if it does, then only at a very slow rate with m.

B. GAUSSIAN LIMIT FOR RELEVANT

The analysis for C_{nr} was possible mainly because the TPs were assumed to be small. For the class C_{r} of relevant documents, this does not have to be the case anymore. Actually, the introduction of the TPs does not seem to make sense anymore if they have to be considered close to 1.

For C_{r}, it seems to be more appropriate to adopt the picture of P_m to be centered around a point $\boldsymbol{q}' \in I\!R^m$. It could, for example, have a multidimensional Gaussian shape around $\boldsymbol{q}'$, or could be nonzero only inside a hyper-ellipsoid around $\boldsymbol{q}'$ and zero outside. In both examples, the distribution is completely defined by $\boldsymbol{q}'$, and a matrix U_m: for the Gaussian case, it is the variance matrix, and the other case the matrix determines the shape of the ellipsoid. We shall assume that the distribution of C_{r} can be defined by these three elements: the center $\boldsymbol{q}'$, the 'shape'-matrix U_m, and a function that determines the rate of decrease (reasonably fast for a Gaussian, infinitely fast for the ellipsoid, and so on). We summarize the above as

ASSUMPTION 4. *in the case of C_{r}, for every m there is an invertible $m \times m$-matrix U_m and a point $\boldsymbol{q}' \in I\!R^m$ such*

that

$$\mathsf{P}_m(d\boldsymbol{\omega}) = \frac{|\det U_m|}{\nu_m}\, f(\,\|U_m(\boldsymbol{\omega}-\boldsymbol{q}')\|^2\,)\, d\boldsymbol{\omega} \ , \qquad (38)$$

where $\nu_m := \int_{I\!R^m} f(\,\|\boldsymbol{\omega}\|^2\,) d\boldsymbol{\omega}$ *is the volume of the function* f *in* $I\!R^m$*, and this function is such that* ν_m *does not grow faster with* m *than a power of* $m!$*.*

The factor $|\det U_m|$ is necessary for the correct normalization of the probability distribution in $I\!R^m$. For example with the Gaussian shape, $(U_m^T U_m)^{-1}$ is the variance matrix in this formulation, and $f(x^2) = \exp(-\frac{1}{2}x^2)$. Notice that P_m induces the 'natural' metric $\|\boldsymbol{\omega}\|_{U_m} := \|U_m\boldsymbol{\omega}\|$.

Let us denote

$$V_m := (U_m^{-1})^T \quad \text{and} \quad \alpha_m := \sqrt{2\pi\,\frac{\nu_{m-2}}{\nu_m}} \ . \qquad (39)$$

Let furthermore S_m be the random variable representing the score of documents from C_{r} with probability measure (38). We will prove that, under the above assumption,

THEOREM 2. *the limiting variable of the sequence*

$$\sigma_m := \alpha_m\,\frac{S_m - \langle \boldsymbol{q}, \boldsymbol{q}'\rangle}{\|V_m\boldsymbol{q}\|}$$

is a normal variable.

Furthermore, we will show that the distribution of σ_m approaches the Gaussian such that corrections go to zero as $1/m$.

We start with expressing $\nu_m := \int_{I\!R^m} f(\,\|\boldsymbol{\omega}\|^2\,)d\boldsymbol{\omega}$ in terms of an 1-dimensional integral. This is possible because the integrand only depends on the length of $\boldsymbol{\omega}$, so that we can go over to spherical coordinates and write

$$\nu_m = \gamma_m \int_0^\infty f(x^2) x^{m-1}\, dx \ , \quad \gamma_m := \frac{2\pi^{\frac{m}{2}}}{\Gamma(\frac{m}{2})} \ , \qquad (41)$$

where γ_m is the volume of an m-dimensional sphere with unit radius, and Γ denotes the gamma function. At this point, we want to note a few facts we shall need later. Firstly, we have

$$\frac{\gamma_{m-1}}{\gamma_{m-3}} = \pi\,\frac{\Gamma(\frac{m-3}{2})}{\Gamma(\frac{m-1}{2})} = \pi\,\frac{\Gamma(\frac{m-3}{2})}{\frac{m-3}{2}\,\Gamma(\frac{m-3}{2})} = \frac{2\pi}{m-3} \ . \qquad (42)$$

Secondly, since we demand that ν_m exists for every m, we obviously have

$$\lim_{x\to\infty} f(x^2) x^{m-1} = 0 \quad \text{for every } m \ . \qquad (43)$$

Thirdly, since we demand that ν_m does not grow faster with m than a power of $m!$, we have

$$\lim_{m\to\infty} \frac{\alpha_{m-2}}{\alpha_m} = \lim_{m\to\infty} \sqrt{\frac{\nu_m}{\nu_{m-2}} \times \frac{\nu_{m-4}}{\nu_{m-2}}} = 1 \ . \qquad (44)$$

We shall prove the theorem by proving that the moments of the the variables σ_m converge towards the moments of a Gaussian variable. Under the distribution (38) the variable σ_m has moments

$$\begin{aligned}\mathsf{E}(\sigma_m^r) &= \int_{I\!R^m} \left(\alpha_m\,\frac{\langle \boldsymbol{q},\boldsymbol{\omega}\rangle - \langle \boldsymbol{q},\boldsymbol{q}'\rangle}{\|V_m\boldsymbol{q}\|}\right)^r \mathsf{P}_m(d\boldsymbol{\omega}) \\ &= \frac{\alpha_m^r}{\nu_m}\int_{I\!R^m} \frac{\langle V_m\boldsymbol{q},\boldsymbol{\omega}\rangle^r}{\|V_m\boldsymbol{q}\|^r} f(\,\|\boldsymbol{\omega}\|^2\,)\, d\boldsymbol{\omega} \ , \end{aligned} \qquad (45)$$

where we performed the substitution $\boldsymbol{\omega} \mapsto U_m^{-1}\boldsymbol{\omega} + \boldsymbol{q}'$ on the integration variable. To prove that all moments exist, we can apply the Schwartz inequality, and go over to spherical coordinates to find that

$$\mathsf{E}(|\sigma_m|^r) \le \frac{\alpha_m^r \gamma_m}{\nu_m}\int_0^\infty x^r f(x^2) x^{m-1}\, dx = \alpha_m^r\,\frac{\nu_{m+r}\gamma_m}{\nu_m\gamma_{m+r}} \ .$$

In order to evaluate (45) further, we note that every $\boldsymbol{\omega}$ can be written as a linear combination of $\boldsymbol{\omega}_{\mathrm{p}}$ parallel to $V_m\boldsymbol{q}$ and an orthogonal component $\boldsymbol{\omega}_{\mathrm{o}}$, so that $\|\boldsymbol{\omega}\|^2 = \|\boldsymbol{\omega}_{\mathrm{p}}\|^2 + \|\boldsymbol{\omega}_{\mathrm{o}}\|^2$. Furthermore, we can always perform an orthogonal basis transformation such that $\boldsymbol{\omega}_{\mathrm{p}}$ lies along a coordinate axis of $I\!R^m$, so we can write $\langle V_m\boldsymbol{q},\boldsymbol{\omega}\rangle = \|V_m\boldsymbol{q}\|\omega_{\mathrm{p}}$, and

$$\mathsf{E}(\sigma_m^r) = \frac{\alpha_m^r}{\nu_m}\int_{-\infty}^\infty y^r \int_{I\!R^{m-1}} f(\,y^2 + \|\boldsymbol{\omega}_{\mathrm{o}}\|^2\,)\, d\boldsymbol{\omega}_{\mathrm{o}}\, dy \ .$$

The integrand is spherical symmetric in $\boldsymbol{\omega}_{\mathrm{o}}$, so that we can go over to spherical coordinates again, and the integral over $I\!R^{m-1}$ reduces to an 1-dimensional integral

$$\mathsf{E}(\sigma_m^r) = \int_{-\infty}^\infty y^r f_m(y)\, dy \ , \qquad (47)$$

where

$$f_m(y) := \frac{\gamma_{m-1}}{\nu_m\alpha_m}\int_0^\infty f\Big(\frac{y^2}{\alpha_m^2} + x^2\Big)\, x^{m-2}\, dx \ . \qquad (48)$$

So the moments of the variable σ_m are equal to the moments of a variable with probability density f_m.

We will now prove that the sequence of density functions f_m has a Gaussian limit. Denote the derivative of f by f', then

$$\begin{aligned}\frac{df_m(y)}{dy} &= \frac{\gamma_{m-1}}{\nu_m\alpha_m}\,\frac{2y}{\alpha_m^2}\int_0^\infty f'\Big(\frac{y^2}{\alpha_m^2} + x^2\Big)\, x^{m-2}\, dx \\ &= -\frac{\gamma_{m-1}}{\nu_m\alpha_m}\,\frac{y}{\alpha_m^2}(m-3)\int_0^\infty f\Big(\frac{y^2}{\alpha_m} + x^2\Big)\, x^{m-4}\, dx \\ &= -y\,\frac{\alpha_{m-2}}{\alpha_m}\, f_{m-2}\Big(\frac{\alpha_{m-2}}{\alpha_m}\, y\Big) \ , \end{aligned}$$

where we applied partial integration and used (43) in the second step, and used (42) and the definition of α_m in the last step. Using (44), we find that the limiting density f_∞ satisfies the differential equation

$$\frac{df_\infty(y)}{dy} = -y f_\infty(y) \ ,$$

which has a Gaussian with zero mean and unit variance as solution. Notice that convergence via the differential equation implies *pointwise* convergence, so that we can conclude that the moments $\mathsf{E}(\sigma_m^r)$ become those of a Gaussian distribution with zero mean and unit variance. This then, leads to the conclusion that σ_m becomes a normal variable.

One might argue that f has to be continuous for this proof, for its derivative is used. This derivative, however, only shows up under an integral, so that it is well defined for discontinuous functions with the help of Dirac distributions.

To answer the question how fast the Gaussian limit appears, we just take $\nu_m = a(m!)^k + \mathcal{O}(\,(m!)^k\,)$ with some $a, k > 0$, so that it is easy to see that

$$\frac{\alpha_{m-2}}{\alpha_m} = 1 + \mathcal{O}\Big(\frac{1}{m}\Big) \ , \qquad (50)$$

and we can conclude that the distribution converges to the Gaussian as $1/m$.