Where to Stop Reading a Ranked List?*

Avi Arampatzis[1] **Jaap Kamps**[1,2]

[1] Archives and Information Studies, Faculty of Humanities, University of Amsterdam
[2] ISLA, Informatics Institute, University of Amsterdam

Abstract: We document our participation in the TREC 2008 Legal Track. This year we focused solely on selecting rank cut-offs for optimizing the given evaluation measure per topic.

1 Introduction

In recall-oriented retrieval setups, such as the Legal Track, ranked retrieval has a particular disadvantage in comparison with traditional Boolean retrieval: there is no clear cut-off point where to stop consulting results. It is expensive to give a ranked list with too many results to litigation support professionals paid by the hour. This may be one of the reasons why ranked retrieval has been adopted very slowly in professional legal search.[1]

The "missing" cut-off remains unnoticed by standard evaluation measures: there is no penalty and only possible gain for padding a run with further results. The TREC 2008 Legal Track addresses this head-on by requiring participants to submit such a cut-off value K per topic where precision and recall are best balanced. This year we focused solely on selecting K for optimizing the given F_1-measure. We believe that this will have the biggest impact on this year's comparative evaluation.

The rest of this paper is organized as follows. The method for determining K is presented in Section 2. It depends on the underlying score distributions of relevant and non-relevant documents, which we elaborate on in Section 3. In Section 4 we describe the parameter estimation methods. In Section 5 we discuss the experimental setup, our official submissions, results, and additional experiments. Finally, we summarize the findings in Section 6.

2 Thresholding a Ranked List

Essentially, the task of selecting K is equivalent to thresholding in binary classification or filtering. Thus, we recruited and adapted a method first appeared in the TREC 2000 Filtering Track, namely, the *score-distributional threshold optimization* (s-d) [2, 3].

2.1 The S-D Threshold Optimization

Let us assume an item collection of size n, and a query for which all items are scored and ranked. Let $P(s|1)$ and $P(s|0)$ be the probability densities of relevant and non-relevant documents as a function of the score s, and $F(s|1)$ and $F(s|0)$ their corresponding *cumulative distribution functions* (cdfs). Let $G_n \in [0, 1]$ be the fraction of relevant documents in the collection of all n documents, also known as *generality*. The total number of relevant documents in the collection is given by

$$R = n\, G_n \tag{1}$$

while the *expected* numbers of relevant and non-relevant documents with scores greater than s are

$$R_+(s) = R\,(1 - F(s|1)) \tag{2}$$
$$N_+(s) = (n - R)\,(1 - F(s|0)) \tag{3}$$

respectively. The expected numbers of the relevant and non-relevant documents with scores $\leq s$ respectively are

$$R_-(s) = R - R_+(s) \tag{4}$$
$$N_-(s) = (n - R) - N_+(s) \tag{5}$$

Let us now assume an effectiveness measure M of the form of a linear combination the document counts of the categories defined by the four combinations of relevance and retrieval status, for example a linear utility [18]. From the property of expectation linearity, the expected value of such a measure would be the same linear combination of the above four expected document numbers. Assuming that the larger the M the better the effectiveness, the optimal score threshold s_θ which maximizes the expected M is

$$s_\theta = \arg\max_s \{M(R_+(s), N_+(s), R_-(s), N_-(s))\} \tag{6}$$

Given n, the only unknowns which have to be estimated are the densities $P(s|1)$ and $P(s|0)$ (or their cdfs), and the generality G_n.

So far, this is a clear theoretical answer to predicting s_θ for linear effectiveness measures. In Section 2.3 we will see how to deal with non-linear measures, as well as, how to predict rank (rather than score) cut-offs.

*The programming code implementing the methods described in this paper will be made publicly available; for information on how to obtain it, please contact the authors.

[1] In fact, to the surprise of many, at the TREC 2007 Legal Track the Boolean reference run outperformed the ranked retrieval models at the rank cut-off of the Boolean set size.

Report Documentation Page	*Form Approved* *OMB No. 0704-0188*

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **NOV 2008**	2. REPORT TYPE	3. DATES COVERED **00-00-2008 to 00-00-2008**

4. TITLE AND SUBTITLE **Where to Stop Reading a Ranked List?**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Amsterdam,Archives and Information Studies,Faculty of Humanities,1012 ZA Amsterdam The Netherlands,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

Seventeenth Text REtrieval Conference (TREC 2008) held in Gaithersburg, Maryland, November 18-21, 2008. The conference was co-sponsored bythe National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA).

14. ABSTRACT

see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**14**	

2.2 Probability Thresholds

Given the two densities and the generality defined earlier, scores can be normalized to probabilities of relevance straightforwardly [2, 14] by using the Bayes' rule.

Normalizing to probabilities is very important in tasks where several rankings need to be fused or merged such as in meta-search/fusion or distributed retrieval. This may also be important for thresholding when documents arrive one by one and decisions have to be made on the spot, depending on the measure under optimization. Nevertheless, it is unnecessary for thresholding rankings since optimal thresholds can be found on their scores directly, and it is furthermore unsuitable given F_1 as the evaluation measure.

While for some measures there exists an optimal *fixed* probability threshold, for others it does not. Lewis [13] formulates this in terms of whether or not a measure satisfies the *probability thresholding principle*, and proves that the F measure does not satisfy it. In other words, how a system should treat documents with, e.g., 50% chance of being relevant depends on how many documents with higher probabilities are available.

The last-cited study also questions whether, for a given measure, an optimal threshold (not necessarily a probability one) exists, and goes on to re-formulate the *probability ranking principle for binary classification*. A theoretical proof is provided about the F measure satisfying the principle, so such an optimal threshold does exist. It is just a different *rank* or *score* threshold for each ranking.

2.3 The S-D Rank Optimization

The s-d threshold optimization method is based on the assumption that the measure M is a linear combination of the document counts of the four categories defined by the user and system decisions about relevance and retrieval status. However, measure linearity is not always the case, e.g. the F measure is non-linear.

Non-linearity complicates the matters in the sense that the expected value of M cannot be easily calculated. Given a ranked list, some approximations can be made simplifying the issue. If G_n, $F(s|1)$, and $F(s|0)$ are estimated on a given ranking, then Equations 2–5 are good approximations of the *actual* document counts. Plugging those counts into M, we can now talk of actual M values rather than expected. The score threshold which maximizes M is given by Equation 6.

While M can be optimal anywhere in the score range, with respect to optimizing rank cutoffs we only have to check its value at the scores corresponding to the ranked documents, plus one extra point to allow for the possibility of an empty optimal retrieved set. Let s_k be the score of the kth ranked document, and define M_k as follows:

$$M_k = \begin{cases} M(R_+(s_k), N_+(s_k), R_-(s_k), N_-(s_k)) & k = 1, \dots, n \\ M(0, 0, R, n - R) & k = 0 \end{cases}$$

The optimal rank K is $\arg\max_k M_k$. This allows for K to become 0, meaning that no document should be retrieved.

3 Score Distributions

Let us now elaborate on the form of the two densities $P(s|1)$ and $P(s|0)$ of Section 2.1 and their estimation. [2]

Score distributions have been modeled since the early years of IR with various known distributions [6, 7, 20, 21]. However, the trend during the last few years, which has started in [3] and followed up in [1, 2, 8, 14, 22], has been to model score distributions by a mixture of normal-exponential densities: normal for relevant, exponential for non-relevant.

Despite its popularity, it was pointed out recently that, under a hypothesis of how systems should score and rank documents, this particular mixture of normal-exponential presents a theoretical anomaly [17]. In practice, nevertheless, it has stand the test of time in the light of

- its (relative) ease to calculate,
- good experimental results, and
- lack of a proven alternative.

The reader should keep in mind that the normal-exponential mixture fits some retrieval models better than others, or it may not fit some data at all. As a rule of thumb, candidates for good fits are scoring functions in the form of a linear combination of query-term weights, e.g. tf.idf, cosine similarity, and some probabilistic models [2]. Also, long queries [2] or good queries/systems [14] seem to help.

In this paper, we do not set out to investigate alternative mixtures. We theoretically extend and refine the current model in order to account for practical situations, deal with its theoretical anomaly, and improve its computation. We also check its goodness-of-fit to empirical data using a statistical test; a check that has not been done before as far as we are concerned. At the same time, we explicitly state all parameters involved, try to minimize their number, and find for them a robust set of values.

3.1 The Normal-Exponential Model

Let us consider a general retrieval model which in theory produces scores in $[s_{\text{min}}, s_{\text{max}}]$, where $s_{\text{min}} \in \mathbb{R} \cup \{-\infty\}$ and $s_{\text{max}} \in \mathbb{R} \cup \{+\infty\}$. By using an exponential distribution, which has semi-infinite support, the applicability of the s-d model is restricted to those retrieval models for which $s_{\text{min}} \in \mathbb{R}$. The two densities are given by

$$P(s|1) = \frac{1}{\sigma}\, \phi\left(\frac{s-\mu}{\sigma}\right) \qquad \sigma > 0,\ \mu, s \in \mathbb{R} \tag{7}$$

$$P(s|0) = \psi(s - s_{\text{min}}; \lambda) \qquad \lambda > 0,\ s \geq s_{\text{min}} \tag{8}$$

where $\phi(.)$ is the density function of the standard normal distribution, i.e. with a mean of 0 and standard deviation of 1,

[2] Probabilistic foundations necessary to follow the discussion can be found in several sources, [e.g., 10, 11, 16]. Where the derivation of a formula is obvious or it can easily be found in the literature, we give directly the result. Otherwise, we show its derivation in Appendix B.

and $\psi(.)$ is the standard exponential density (Equations 18–19 in Appendix B). The corresponding cdfs are given by Equations 20 and 22. The total score distribution is written as

$$P(s) = (1 - G_n)\, P(s|0) + G_n P(s|1)$$

where $G_n \in [0, 1]$. Hence, there are 4 parameters to estimate, λ, μ, σ, and G_n.

3.2 Problems of the Normal-Exponential Model

Over the years, two main problems of the normal-exponential model have been identified. We describe each one of them, and then introduce new models which eliminate the first problem and deal partly with the other.

3.2.1 Support Incompatibility

Although we already generalized somewhat above by introducing a *shifted exponential*, the mix, as it has been used in all related literature so far, has a support incompatibility problem: while the exponential is defined at or above some $s_{\min}$, the normal has a full real axis support. This is a theoretical problem which is solved by the new models we will introduce.

3.2.2 Recall-Fallout non-Convexity

From the point of view of how scores or rankings of IR systems should be, Robertson [17] formulates the recall-fallout convexity hypothesis:

> *For all good systems, the recall-fallout curve (as seen from [. . .] recall=1, fallout=0) is convex.*

Similar hypotheses can be formulated as a conditions on other measures, e.g., the probability of relevance should be monotonically increasing with the score; the same should hold for *smoothed* precision. Although, in reality, these conditions may not always be satisfied, they are expected to hold for good systems, i.e. those producing rankings satisfying the *probability ranking principle* (PRP), because their failure implies that systems can be easily improved.

As an example, let us consider smoothed precision. If it declines as score increases for a part of the score range, that part of the ranking can be improved by a simple random re-ordering [19]. This is equivalent of "forcing" the two underlying distributions to be uniform (i.e. have linearly increasing cdfs) in that score range. This will replace the offending part of the precision curve with a flat one—the least that can be done— improving the overall effectiveness of the system.

Such hypotheses put restrictions on the relative forms of the two underlying distributions. The normal-exponential mixture violates such conditions, only (and always) at both ends of the score range. Although the low-end scores are of insignificant importance, the top of the ranking is very significant, especially for low R topics. The problem is a manifestation of the fact that an exponential tail extends further than a normal one.

To complicate matters further, our data suggest that such conditions are violated at a different score s_c for the probability of relevance and for precision. Since the F-measure we are interested in is a combination of recall and precision (and recall by definition cannot have a similar problem), we find s_c for precision. We force the distributions to comply with the hypothesis only when $s_c < s_1$, where s_1 the score of the top document; otherwise, the theoretical anomaly does not affect the score range. If $s_{\max}$ is finite, then two uniform distributions can be used in $[s_c, s_{\max}]$ as mentioned earlier. Alternatively, preserving a theoretical support in $[s_{\min}, +\infty)$, the relevant documents distribution can be forced to an exponential in $[s_c, +\infty)$ with the same λ as this of the non-relevant. We apply the alternative.

In fact, rankings can be further improved by reversing the offending sub-rankings; this will force the precision to increase with an increasing score, leading to better effectiveness than randomly re-ordering the sub-ranking. However, the big question here is whether the initial ranking satisfies the PRP or not. If it does, then the problem is an artifact of the normal-exponential model and reversing the sub-ranking may actually be dangerous to performance. If it does not, then the problem is inherent in the scoring formula producing the ranking. In the latter case, the normal-exponential model cannot be theoretically rejected, and it may even be used to detect the anomaly and improve rankings.

It is difficult, however, to determine whether a single ranking satisfies the PRP or not; it is well-known since the early IR years that precision for single queries is erratic, especially at early ranks, justifying the use of interpolated precision. On the one hand, according to interpolated precision all rankings satisfy the PRP, but this is *forced* by the interpolation. On the other hand, according to simple precision some of our rankings do not *seem* to satisfy the PRP, but we cannot determine this for sure. We would expect, however, that using precision averaged over all topics should produce a—more or less—declining curve with an increasing rank. Figure 1 suggests that the off-the-shelf system we currently use produces rankings that may not satisfy the PRP for ranks 5,000 to 10,000, on average.

Consequently, we rather leave open the question of whether the problem is inherent in some scoring functions or introduced by the combined use of normal and exponential distributions. Being conservative, we just randomize the offending sub-rankings rather than reversing them. The impact of this on thresholding is that the s-d method turns "blind" inside the upper offending range; as one goes down the corresponding ranks, precision would be flat, recall naturally rising, so the optimal F_1 threshold can only be below the range.

We will use new models that, although they do not eliminate the problem, also do not always violate such conditions imposed by the PRP (irrespective of whether it holds or not).

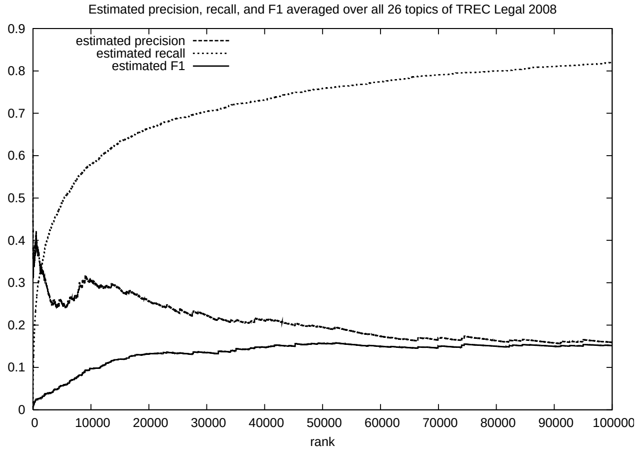


Figure 1: Precision, Recall, and F_1—as these are estimated by TREC's deep-sampling method—averaged over all 26 topics of TREC Legal 2008. By rank 100,000, precision is still flat rather than declining, recall is still rising, so F_1 has not yet peaked; this suggests that there are optimal K's larger than 100,000. Systems correctly predicting K's larger than 100,000 do not get credit.

3.3 The Truncated Normal-Exponential Model

In order to enforce support compatibility, Arampatzis et al. [5] introduced truncated models which we will discuss in this and the next section. They introduced a left-truncated at $s_{\min}$ normal distribution for $P(s|1)$. With this modification, we reach a new mixture model for score distributions with a semi-infinite support in $[s_{\min}, +\infty)$, $s_{\min} \in \mathbb{R}$.

In practice, however, scores may be naturally bounded (by the retrieval model) or truncated to the upside as well. For example, cosine similarity scores are naturally bounded at 1. Scores from probabilistic models with a (theoretical) support in $(-\infty, +\infty)$ are usually mapped to the bounded $(0, 1)$ via a logistic function. Other retrieval models may just truncate at some maximum number for practical reasons. Consequently, it makes sense to introduce a right-truncation as well, for both the normal and exponential densities.

Depending on how one wants to treat the leftovers due to the truncations, two new models may be considered.

3.3.1 Theoretical Truncation

There are no leftovers (Figure 2). The underlying theoretical densities are assumed to be the truncated ones, normalized accordingly to integrate to one:

$$P(s|1) = \frac{\frac{1}{\sigma}\,\phi\left(\frac{s-\mu}{\sigma}\right)}{\Phi(\beta) - \Phi(\alpha)} \qquad s \in [s_{\min}, s_{\max}] \tag{9}$$

$$P(s|0) = \frac{\psi(s - s_{\min}; \lambda)}{\Psi(s_{\max} - s_{\min}; \lambda)} \qquad s \in [s_{\min}, s_{\max}] \tag{10}$$

where

$$\alpha = \frac{s_{\min} - \mu}{\sigma} \qquad \beta = \frac{s_{\max} - \mu}{\sigma} \tag{11}$$

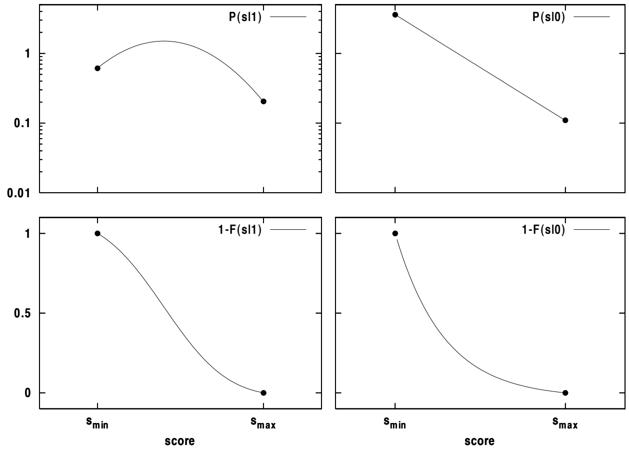


Figure 2: Theoretical truncation.

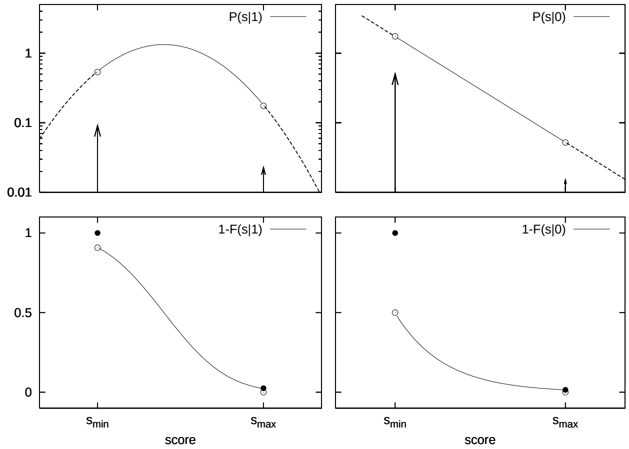


Figure 3: Technical truncation.

$\Phi(.)$ and $\Psi(.)$ are the cdfs of $\phi(.)$ and $\psi(.)$ respectively (Equations 20 and 22). The cdfs of the above $P(s|1)$ and $P(s|0)$ are given by Equations 21 and 23, respectively.

Let S_{rel} and S_{nrel} be the random variables corresponding to the relevant and non-relevant document scores respectively. The expected value and variance of S_{rel} are given by Equations 24 and 25 in Appendix B.3. For S_{nrel} , the corresponding Equations are 26 and 27 in Appendix B.4.

3.3.2 Technical Truncation

The underlying theoretical densities are not truncated, but the truncation is of a "technical" nature. The leftovers are accumulated at the two truncation points introducing discontinuities (Figure 3). For the normal, the leftovers can easily be calculated:

$$P(s|1) = \begin{cases} \Phi(\alpha)\,\delta(s - s_{\min}) & s = s_{\min} \\ \frac{1}{\sigma}\,\phi\left(\frac{s-\mu}{\sigma}\right) & s \in (s_{\min}, s_{\max}) \\ (1 - \Phi(\beta))\;\delta(s - s_{\max}) & s = s_{\max} \end{cases}$$

where $\delta(.)$ is Dirac's delta function. For the exponential, while the leftovers at the right side are determined by the right truncation, in order to calculate the ones at the left side requires to assume that the exponential extends below s_{min} to some new minimum score s'_{min}:

$$P(s|0) = \begin{cases} \Psi(s_{\text{min}} - s'_{\text{min}}; \lambda)\, \delta(s - s_{\text{min}}) & s = s_{\text{min}} \\ \psi(s - s'_{\text{min}}; \lambda) & s \in (s_{\text{min}}, s_{\text{max}}) \\ (1 - \Psi(s_{\text{max}} - s'_{\text{min}}; \lambda))\, \delta(s - s_{\text{max}}) & s = s_{\text{max}} \end{cases}$$

The cdfs corresponding to the above densities are:

$$F(s|1) = \begin{cases} \Phi(\frac{s-\mu}{\sigma}) & s \in [s_{\text{min}}, s_{\text{max}}) \\ 1 & s = s_{\text{max}} \end{cases}$$

$$F(s|0) = \begin{cases} \Psi(s - s'_{\text{min}}; \lambda) & s \in [s_{\text{min}}, s_{\text{max}}) \\ 1 & s = s_{\text{max}} \end{cases}$$

The equations in this section simplify somewhat when estimating their parameters from down-truncated ranked lists, as we will see in Section 4.1. We do not need to calculate s'_{min}. If, for some measure, the number of non-relevant documents is required, it can simply be estimated as $n - R$.

The expected values and variances of S_{rel} and S_{nrel}, if needed, have to be calculated starting from Equations 24–27 and taking into account the contribution of the discontinuities. We do not give the formulas in this paper.

3.3.3 The Relation Between the Truncated Models

For both models the right truncation is optional. For $s_{\text{max}} = +\infty$, we get $\Phi(\beta) = \Psi(s_{\text{max}} - s'_{\text{min}}; \lambda) = 1$, leading to left-truncated models; this accommodates retrieval models with scoring support in $[s_{\text{min}}, +\infty)$, $s_{\text{min}} \in \mathbb{R}$. This is the maximum range that can be achieved with the current mixture, since the restriction of a finite s_{min} is imposed by the use of the exponential.

When $s_{\text{min}} \ll \mu \ll s_{\text{max}}$ then $\Phi(\alpha) \approx 0$ and $\Phi(\beta) \approx 1$. If additionally $s'_{\text{min}} = s_{\text{min}}$, then $\Psi(s_{\text{min}} - s'_{\text{min}}; \lambda) = 0$ and $\Psi(s_{\text{max}} - s'_{\text{min}}; \lambda) \approx 1$. Thus we can well-approximate the standard normal-exponential model. Consequently, using a truncated model is a valid choice even when truncations are insignificant.

From a theoretical point of view, it may be difficult to imagine a process producing a truncated normal *directly*. Truncated normal distributions are usually the results of censoring, meaning that the out-truncated data do actually exist. In this view, the technically truncated model may correspond better to the IR reality. This is also in line with the theoretical arguments for the existence of a full normal distribution [2].

Concerning convexity, both truncated models do not always violate such conditions. Consider the problem at the top score range $(s_c, +\infty)$. In the cases of $s_c \geq s_{\text{max}}$, the problem is out-truncated in both models, while—in theory—it still always exists in the original model. The improvement so far is of a rather theoretical nature. In practise, we should be interested in what happens when $s_c < s_1$. Our extended experiments (not reported in this paper) suggest that truncation helps estimation in producing higher numbers of convex fits within the observed score range. Consequently, the benefits are also practical.

These improvements make the original model more general, and it indeed produces better fits on our data. In fact, the truncated distributions should have been used in the past during parameter estimation even for the original normal-exponential model due to down-truncated rankings.

4 Parameter Estimation

The normal-exponential mixture has worked best under the availability of some relevance judgments which serve as an indication about the form of the component densities [3, 8, 22]. In filtering or classification, usually some training data—although often biased—are available. In the current task, however, no relevance information is available.

A method was introduced in the context of fusion which recovers the component densities without any relevance judgments using the Expectation Maximization (EM) algorithm [14]. In order to deal with the biased training data in filtering, the EM method was also later adapted and applied for thresholding tasks [1].[3] Nevertheless, EM was found to be "messy" and sensitive to its initial parameter settings [1, 14]. We will improve upon this estimation method in Section 4.3.

4.1 Down-truncated Rankings

For practical reasons, rankings are usually truncated at some rank $t < n$. Even what is usually considered a full ranking is in fact a collection's subset of those documents with at least one matching term with the query.

This fact has been largely ignored in all previous research using the standard model, despite that it may affect greatly the estimation. For example, in TREC Legal 2007 and 2008, t was $25,000$ and $100,000$ respectively. This results to a left-truncation of $P(s|1)$ which at least in the case of the 2007 data is significant. For 2007 it was estimated that there were more than $25,000$ relevant documents for 13 of the 43 Ad Hoc topics (to a high of more than $77,000$) and the median system was still achieving 0.1 precision at ranks of $20,000$ to $25,000$.

Additionally, considering that the exponential may not be a good model for the whole distribution of the non-relevant scores but only for their high end, some imposed truncation may help achieve better fits. Consequently, all estimations should take place at the top of the ranking, and then get extrapolated to the whole collection. The truncated models of [5] require changes in the estimation formulas.

Let us assume that the truncation score is s_t. For both truncated models, we we need to estimate a two-side truncated normal at s_t and s_{max}, and a shifted exponential by s_t

[3] Another method for producing unbiased estimators in filtering can be found in [22], but it requires relevance judgements.

right-truncated at $s_{\max}$, with $s_{\max}$ possibly be $+\infty$. Thus, the formulas that should be used are Equations 9 and 10 but for α_t instead of α

$$\alpha_t = \frac{s_t - \mu}{\sigma}$$

and for s_t instead of $s_{\min}$. Beyond this, the models differentiate in the way R is calculated.

If G_t is the fraction of relevant documents in the truncated ranking, extrapolating the truncated normal outside its estimation range and appropriately per model in order to account for the remaining relevant documents, the R is calculated as:

- theoretically truncated normal-exponential

$$R = t\,G_t\,\frac{\Phi(\beta) - \Phi(\alpha)}{\Phi(\beta) - \Phi(\alpha_t)}$$

- technically truncated normal-exponential

$$R = t\,G_t\,\frac{1}{\Phi(\beta) - \Phi(\alpha_t)}$$

Consequently, Equation 1 must be replaced by one of the above depending on the model in use, Equations 2 and 3 must be re-written as

$$R_+(s) = t\,G_t\,\left(1 - F(s|1)\right)$$

$$N_+(s) = t\,\left(1 - G_t\right)\left(1 - F(s|0)\right)$$

while Equations 4 and 5 remain the same. $F(s|1)$ and $F(s|0)$ are now the cdfs either of Section 3.3.1 or 3.3.2, depending on which model is used.

In estimating the technically truncated model, if there are any scores equal to $s_{\max}$ or $s_{\min}$ they should be removed from the data-set; these belong to the discontinuous legs of the densities given in Section 3.3.2. In this case, t should be decremented accordingly. In practise, while scores equal to $s_{\min}$ should not exist in the top-t due to the down-truncation, some $s_{\max}$ scores may very well be in the data. Removing these during estimation is a simplifying approximation with an insignificant impact when the relevant documents are many and the bulk of their score distribution is below $s_{\max}$, as it is the case in current experimental setup. As we will see next, while we do not use the $s_{\max}$ scores during fitting, we take them into account during goodness-of-fit testing; using multiple such fitting/testing rounds, this reduces the impact of the approximation.

4.2 Score Preprocessing

Our scores have a resolution of 10^{-6}. Obviously, LUCENE rounds or truncates the output scores, destroying information. In order to smooth out the effect of rounding in the data, we add $\Delta s = \mathrm{rand}(10^{-6}) - 0.5 * 10^{-6}$ to each datum point, where $\mathrm{rand}(x)$ returns a uniformly-distributed real random number in $[0, x)$.

Beyond using all scores available and in order to speed up the calculations, we also tried stratified down-sampling to keep only 1 out of 2, 3, or 10 scores.[4] Before any down-sampling, all datum points were smoothed by replacing them with their average value in a surrounding window of 2, 3, or 10 points, respectively.

In order to obtain better exponential fits we may further left-truncate the rankings at the mode of the observed distribution. We bin the scores (as described in Section A.1), find the bin with the most scores, and if that is not the leftmost bin then we remove all scores in previous bins.

4.3 Expectation Maximization

EM is an iterative procedure which converges locally [9]. Finding a global fit depends largely on the initial settings of the parameters.

4.3.1 Initialization

We tried numerous initial settings, but no setting seemed universal. While some settings helped a lot some fits, they had a negative impact on others. Without any indication of the form, location, and weighting of the component densities, the best fits overall were obtained for randomized initial values, preserving also the generality of the approach:[5]

$$G_{t,\mathrm{init}} = \mathrm{rand}(1)\,, \qquad \lambda_{\mathrm{init}} = \max(\epsilon, \mathrm{rand}(\mu_s - s_t))^{-1}$$

$$\mu_{\mathrm{init}} = s_{\min} + \mathrm{rand}(s_1 - s_{\min})$$

$$\sigma^2_{\mathrm{init}} = \max(\epsilon^2, (1 + c_1 \mathrm{rand}(1))^2 \sigma_s^2 - \lambda_{\mathrm{init}}^{-2})$$

where s_1 is the maximum score datum, μ_s and σ_s^2 are respectively the mean and variance of the score data, ϵ is an arbitrary small constant which we set equal to the width of the bins (see Appendix A.1), and $c_1 \in (0, +\infty)$ is another constant which we explain below.

Assuming that no information is available about the expected R, not much can be done for $G_{t,\mathrm{init}}$, so it is randomized using its whole supported range. Next we assume that right-truncation of the exponential is insignificant, which seems to be the case in our current experimental set-up.

If there are no relevant documents, then $\mu_s - (s_t - s_{\min}) \approx \lambda^{-1} + s_{\min}$. From the last equation we deduce the minimum λ_{init}. Although in general, there is no reason why the exponential cannot fall slower that this, from an IR perspective it should not, or $\mathrm{E}(S_{\mathrm{nrel}})$ would get higher than $\mathrm{E}(S_{\mathrm{rel}})$.

The μ_{init} given is suitable for a full normal, and its range should be expanded in both sides for a truncated one because

[4] In order not to complicate things further, we do not include the down-sampling into the formulas in this paper; it is not difficult to see where things should be weighted inversely proportional to the sampling probability.

[5] With some (even biased) training data, suitable initial parameter settings are given in [1]. Without any training data, assuming that the relevant documents are much fewer than non-relevant by rank t, initial parameters can be estimated as described in [14]; unfortunately this assumption cannot be made in TREC Legal due to the large variance of estimated R and topics with $R > t$.

the mean of the corresponding full normal can be below $s_{\min}$ or above s_1. Further, μ_{init} can be restricted based on the hypothesis that for good systems should hold that $\mathrm{E}(S_{\text{rel}}) > \mathrm{E}(S_{\text{nrel}})$. We have not worked out these improvements.

The variance of the initial exponential is $\lambda_{\text{init}}^{-2}$. Assuming that the random variables corresponding to the normal and exponential are uncorrelated, the variance of the normal is $\geq \sigma_s^2 - \lambda_{\text{init}}^{-2}$ which, depending on how λ is initialized, could take values ≤ 0. To avoid this, we take the max with the constant. For an insignificantly truncated normal, $c_1 \approx 0$, while in general $c_1 > 0$, because the variance of the corresponding full normal is larger than what is observed in the truncated data. We set $c_1 = 2$, however, we found its usable range to be $[0.25, 5]$.

4.3.2 Update Equations

For $t \leq n$ observed scores $s_1 \ldots s_t$, and neither truncated nor shifted normal and exponential densities (i.e. for the original model), the update equations are

$$G_{t,\text{new}} = \frac{\sum_i P_{\text{old}}(1|s_i)}{t} \qquad \lambda_{\text{new}} = \frac{\sum_i P_{\text{old}}(0|s_i)}{\sum_i P_{\text{old}}(0|s_i) s_i}$$

$$\mu_{\text{new}} = \frac{\sum_i P_{\text{old}}(1|s_i) s_i}{\sum_i P_{\text{old}}(1|s_i)} \quad \sigma_{\text{new}}^2 = \frac{\sum_i P_{\text{old}}(1|s_i)(s_i - \mu_{\text{new}})^2}{\sum_i P_{\text{old}}(1|s_i)}$$

$P(j|s)$ is given by Bayes' rule $P(j|s) = P(s|j)P(j)/P(s)$, $P(1) = G_t$, $P(0) = 1 - G_t$, and $P(s)$ by Equation 3.1.

We initialize those equations as described above, and iterate them until the absolute differences between the old and new values for μ, λ^{-1}, and $\sqrt{\sigma}$ are all less than .001 $(s_1 - s_{\min})$, and $|G_{t,\text{new}} - G_{t,\text{old}}| < .001$. Like this we target an accuracy of 0.1% for scores and 1 in a 1,000 for documents. We also tried a target accuracy of 0.5% and 5 in 1,000, but it did not seem sufficient.

4.3.3 Correcting for Truncation

If we use the truncated densities (Equations 9 and 10) in the above update equations, the μ_{new} and σ_{new}^2 calculated at each iteration would be the expected value and variance of the truncated normal, not the μ and σ^2 we are looking for. Similarly, $1/\lambda_{\text{new}} + s_t$ would be equal to the expected value of the shifted truncated exponential. Instead of looking for new EM equations, we rather correct to the right values using simple approximations.

Using Equation 26, at the end of each iteration we correct the calculated λ_{new} as

$$\lambda_{\text{new}} \leftarrow \left(\frac{1}{\lambda_{\text{new}}} + s_t + \frac{s_{\max} \exp(-\lambda_{\text{old}}(s_{\max} - s_t)) - s_t}{\Psi(s_{\max} - s_t; \lambda_{\text{old}})} \right)^{-1} \tag{12}$$

using the λ_{old} from the previous iteration as an approximation. Similarly, based on Equations 24 and 25, we correct the calculated μ_{new} and σ_{new}^2 as

$$\mu_{\text{new}} \leftarrow \mu_{\text{new}} - \frac{\phi(\alpha') - \phi(\beta')}{\Phi(\beta') - \Phi(\alpha')} \sigma_{\text{old}} \tag{13}$$

$$\sigma_{\text{new}}^2 \leftarrow \sigma_{\text{new}}^2 \left[1 + \frac{\alpha' \, \phi(\alpha') - \beta' \, \phi(\beta')}{\Phi(\beta') - \Phi(\alpha')} - \left(\frac{\phi(\alpha') - \phi(\beta')}{\Phi(\beta') - \Phi(\alpha')} \right)^2 \right]^{-1} \tag{14}$$

where

$$\alpha' = \frac{s_t - \mu_{\text{old}}}{\sqrt{\sigma_{\text{old}}^2}} \qquad \beta' = \frac{s_{\max} - \mu_{\text{old}}}{\sqrt{\sigma_{\text{old}}^2}}$$

again using the values from the previous iteration.

These simple approximations work, but sometimes they seem to increase the number of iterations needed for convergence, depending on the accuracy targeted. Rarely, and for high accuracies only, the approximations possibly handicap EM convergence; the intended accuracy is not reached for up to 1,000 iterations. Generally, convergence happens in 10 to 50 iterations depending on the number of scores (more data, slower convergence), and even with the approximation EM produces considerably better fits than when using the non-truncated densities. To avoid getting locked in a non-converging loop, despite its rarity, we cap the number of iterations to 100. The end-differences we have seen between the observed and expected numbers of documents due to these approximations have always been less than 4 in 100,000.

4.3.4 Multiple Runs

We initialize and run EM as described above. After EM stops, we apply the χ^2 goodness-of-fit test for the observed data and the recovered mixture (see Appendix A). If the null hypothesis H_0 is rejected, we randomize again the initial values and repeat EM for up to 100 times or until H_0 cannot be rejected. If H_0 is rejected in all 100 runs, we just keep the best fit found. We run EM at least 10 times, even if we cannot reject H_0 earlier. Perhaps a maximum of 100 EM runs is an overkill, but we found that there is significant sensitivity to initial conditions.

4.3.5 Rejecting Fits on IR Grounds

Some fits, irrespective of their quality, can be rejected on IR grounds. Firstly, it should hold that $R \leq n$, however, since each fit corresponds to $t\ (1 - G_t)$ non-relevant documents, we can tighten the inequality somewhat to:

$$R \leq n - t\ (1 - G_t) \tag{15}$$

This is a very light condition, which should handle a few extremities. Secondly, concerning the random variables S_{rel} and S_{nrel}, one would expect:

$$\mathrm{E}(S_{\text{rel}}) > \mathrm{E}(S_{\text{nrel}}) \tag{16}$$

This is rather only a hypothesis—not a requirement—that good systems should satisfy and there are no guarantees. We have not been able so far to motivate any inequality on score variances.

We are still experimenting with such conditions, and we have not applied them for producing any of the end-results reported in this paper.

Table 1: The effects of sampling and binning on fitting quality, and convexity of fits.

run	$\widetilde{M}$	$M > 190$	H_0 no reject	$k_c > 1$	$\widetilde{k_c}$	$k_c > \widetilde{R}$	comments
2007-default	56.5	4 (8%)	2 (4%)	33 (66%)	29	5 (10%)	no smth or sampling
2007-A	37	1 (2%)	32 (64%)	40 (80%)	34	5 (10%)	smth + 1/3 strat. sampl.
2007-B	36	1 (2%)	30 (60%)	32 (64%)	61.5	1 (2%)	smth + 1/3 strat. sampl.
2008-default	93	6 (13%)	0 (0%)	29 (64%)	89	0 (0%)	no smth or sampling
2008-A	63	1 (2%)	5 (11%)	30 (67%)	98	0 (0%)	smth + 1/3 strat. sampl.
2008-B	66	4 (9%)	9 (20%)	31 (69%)	45	1 (2%)	smth +1/3 strat. sampl.

4.4 Fitting Results and Analysis

While the s-d method is non-parametric, there are several parameters in recovering the mixture of the densities: smoothing and sampling (both optional), binning, EM initialization and targeted accuracy, rejection conditions, and maybe others. Table 1 provides some data on the fits resulting from the above procedure. The *default* and *A* runs use the theoretical truncation of Section 3.3.1; the *B* runs use the technical truncation of Section 3.3.2.

4.4.1 Sampling, Binning, and Quality of the Fits

Down-sampling has the effect of eliminating some of the right tails, leading to fewer bins when binning the data. Moreover, the fewer the scores, the less EM iterations and runs are needed for a good fit (data not shown). Down-sampling the scores helps supporting the H_0. At 1 out of 3 stratified sampling, the H_0 cannot be rejected at a significance level of 0.05 for 60-64% of the 2007 topics and for 20% for the 2008 topics. Non-stratified down-sampling with 0.1 probability raises this to 42% for the 2008 topics. Extreme down-sampling to keep only around 1,000 to 5,000 scores supports the H_0 in almost all fits.

Consequently, the number of scores and bins plays a big role in the quality of the fits according to the χ^2 test; there is a positive correlation between the median number of bins $\widetilde{M}$ and the percentage of rejected H_0. This effect does not seem to be the result of information loss due to down-sampling; we still get more support for the H_0 when reducing the number of scores by down-truncating the rankings instead of down-sampling them. This is an unexpected result; we rather expected that the increased number of scores and bins is dealt with by the increased degrees of freedom parameter of the corresponding χ^2 distributions. Irrespective of sampling and binning, however, all fits look reasonably well to the eye.

4.4.2 A Score Continuity Problem?

In all runs, for a small fraction of topics (2-13%) the optimum number of bins M is near ($< 5\%$ difference) to our capped value of 200. For most of these topics, when looking for the optimal number of bins in the range $[5, 1000]$ (numbers are tried with a step of 5%) the binning method does not converge. This means there is no optimal binning as the algorithm identifies the discrete structure of data as being a more salient feature than the overall shape of the density function. Figure 4 demonstrates this.

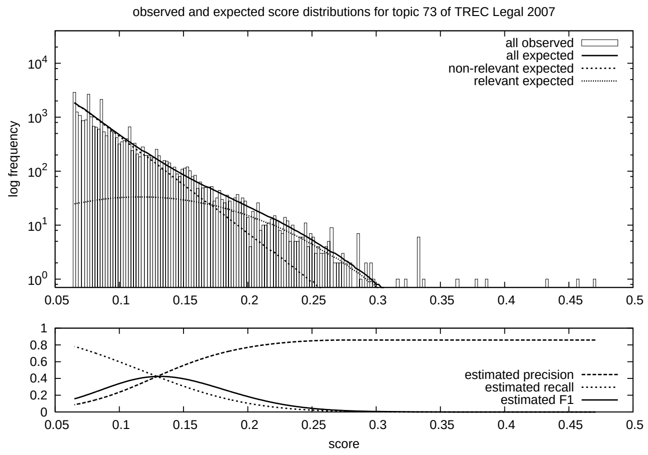


Figure 4: The optimal number of bins does not seem to converge, so it is capped at 200. Due to the high number of bins, the best fit found has a large $\chi^2 = 2231.4$. Combining bins with expected frequency < 5 on the right tail, minus 4 the parameters we estimate, gives 84 degrees of freedom for the χ^2 distribution and a critical value of 106.4 at .05 significance. The upper-probability of the fit is practically 0, nevertheless, it looks reasonably well to the eye.

Since the scores are already randomized to account for rounding (Section 4.2), the discrete structure of the data is not a result of rounding but it rather comes from the retrieval model itself. Internal statistics are usually based on document and word counts; when these are low, statistics are "rough", introducing the discretization effect.

4.4.3 Convexity of Fits

Concerning the theoretical anomaly of the normal-exponential mixture, we investigate the number of fits presenting the anomaly within the observed score range, i.e. at a rank below rank-1 ($k_c > 1$).[6] We see that the anomaly shows up in a large number of topics (64-80%). The impact of non-convexity on the s-d method is that the method turns "blind" at rank numbers $< k_c$ restricting the estimated op-

[6] In our context we re-formulated the recall-fallout convexity hypothesis as a condition on smoothed precision. So there is no issue of convexity but rather the issue of the precision monotonically declining with the score. However, we stick to using the term "convexity" in describing the problem.

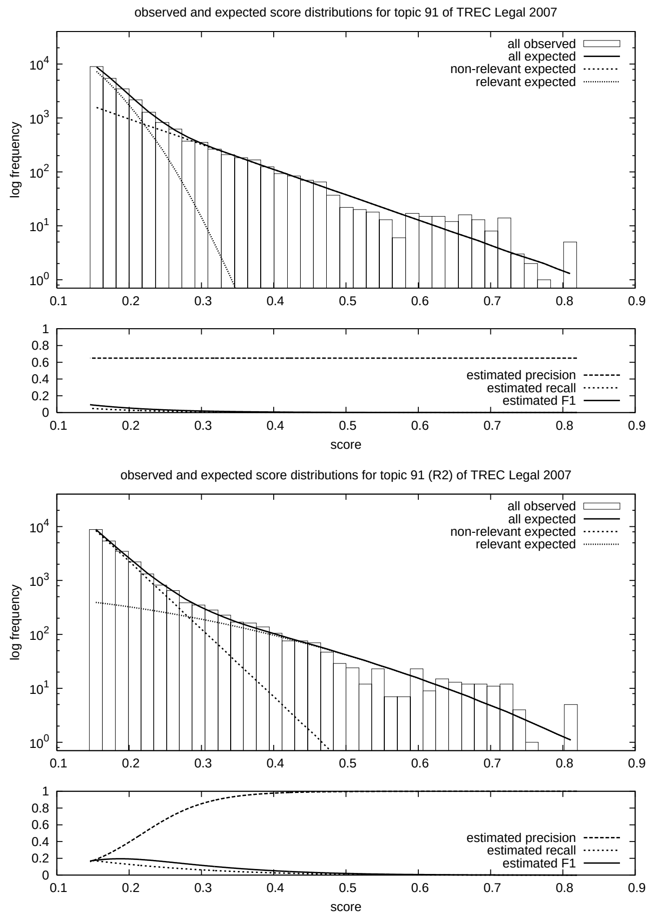


Figure 5: For topic 91 (top plot), the fit looks good but has a convexity problem in the whole ranking ($k_c \geq 25,000$), indicated by having to flatten its precision in the whole range. Alternatively, the fit could have been rejected on IR grounds. By enabling the condition of Equation 16, i.e. the expected relevant score should be larger than the expected non-relevant, the method would have rejected the fit and produce another one (bottom plot) with a slightly larger χ^2 but no convexity problem. (Both datasets are downsampled; the slight variation of the observed data across the plots are due to different samples used.)

timal thresholds wtih $K \geq k_c$. However, the median rank number $\widetilde{k_c}$ down to which the problem exists is very low compared to the median estimated number of relevant documents $\widetilde{R}$ (7,484 or 32,233), so $K < k_c$ is unlikely on average anyway and thresholding should not be affected. Consequently, the data suggest that the non-convexity should have an insignificant impact on s-d thresholding.

For a small number of topics (0-10%), the problem appears for $k_c > \widetilde{R}$ and non-convexity should have a significant impact. Still, we argue that for a good fraction of such topics, a large k_c indicates a fitting problem rather than a theoretical one. Figure 5 explains this further.

4.4.4 Ranking-length Bias

Since there are more data at lower scores, EM results in parameter estimates that fit the low scores better than the high scores. This is exactly the opposite of what is needed for IR purposes, where the top of rankings is more important. It also introduces a weak but undesirable bias: the longer the input ranked list, the lower the estimates of the means of the normal and exponential; this usually results in larger estimations of R and K.

Trying to neutralize the bias without actually removing it, input ranking lengths can better be chosen according to the expected R. This also makes sense for the estimation method irrespective of biases: we should not expect much when trying to estimate, e.g., an R of 100,000 from only the top-1000. As a rule-of-thumb, we recommend input ranking lengths of around 1.5 times the expected R with a minimum of 200. According to this recommendation, the 2007 rankings truncated at 25,000 are spot on, but the 100,000 rankings of 2008 are falling short by 20%.

4.5 Summary and Future Improvements

Recovering the mixture with EM has been proven to be "tricky". However, with the improvements presented in this paper, we have reached a rather stable behavior which produces usable fits.

EM's initial parameter settings can further be tightened resulting in better estimates in less iterations and runs, but we have rather been conservative in order to preserve the generality.

As a result of how EM works—giving all data equal importance—a weak but undesirable ranking-length bias is present: the longer the input ranking, the larger the R estimates. Although the problem can for now be neutralized by choosing the input lengths in accordance with the expected R, any future improvements of the estimation method should take into account that score data are of unequal importance: data should be fitted better at their high end.

Whatever the estimation method, conditions for rejecting fits on IR grounds such as those investigated in Section 4.3.5, seem to have a potential for further considerable improvements.

5 Experiments

In this section, we will conduct a range of experiments with the truncated models of [5], which we discussed in great detail above. Since our focus is the thresholding problem, we use an off-the-shelf retrieval system: the vector-space model of Apache's `Lucene`.

More information about the collection, topics, and evaluation measures can be found in the overview paper in this volume, and at the TREC Legal web-site.

5.1 Runs

For TREC Legal 2007 and 2008 we created the following runs:

Legal07 Off-the-shelf LUCENE using the `RequestText` as query, on a stemmed index, using the generic SMART stoplist. The 2007 rankings are truncated at 25k results.

This run is the run labeled `catchup0701t` in [4].

Legal08 Same as above, but in pre-processing this year's topics, we used the `RequestText` field stop-listed by an extended list in which we manually included low-content words based on the topics of 2006 and 2007. All 2008 rankings are truncated at 100k items.

This runs is the basis for the official submissions labeled `uva-xcons`, `uva-xb`, and `uva-xk`.

For the threshold optimization, we first apply the original version of the score-distributional threshold optimization as it has been used, for example, in the Filtering track [2, 3]:

sd original First fitting a mixture of normal (for relevant) and exponential (for non-relevant) to the score distribution, and then calculate the rank that maximizes the F_1 measure. Note that the fit may indicate an optimal rank threshold beyond the run's length (25k in 2007 and 100k in 2008), in which case we simply select the final rank.

This run corresponds to our official submission labeled `uva-xk`.

In this paper, we presented an improved version of the sd method in Section 3. The improvements that have the greatest impact on end-user effectiveness are:

1. Use of truncated distributions [5] to account for natural score bounds or truncations.
2. EM is run with different initial parameters, and better termination methods. We also now run it up to 100 times instead of 10.
3. We used the square error before to select the best fit; we replaced this with the χ^2 which is more suitable for distributions.
4. Optimal binning. Before, we used a fixed number of $\max(5, t/200)$ bins, which gave 500 bins (or a bit less after a left-truncation of the data) for the 2008 rankings.

Consequently, we provide here additional runs:

Theoretical Truncation Runs using the *theoretical truncation* of Section 3.3.1. The **B** runs is down-sampled (a stratified sample of 1/3).

Theoretical Truncation Runs using the *technical truncation* of Section 3.3.2. The **A** runs are down-sampled (a stratified sample of 1/3). Details of the effect of sampling and binning on the fits are in Table 1.

Table 2: Ranking quality for the Legal 2007 & 2008. The highest, lowest, and median are of the 23 submissions in 2008 using the `RequestText` field only.

Run	$Prec@5$	$Recall@B$	$F_1@R$
Legal07	0.3302	0.1548	0.1328
Legal08	0.4846	0.2036	0.1709
highest	0.5923	0.2779	0.2173
median	0.4154	0.2036	0.1709
lowest	0.0538	0.0729	0.0694

Table 3: Estimating cut-off K for the Legal 2007 & 2008. The highest, lowest, and median are of the 23 submissions using the `RequestText` field. Statistical significance (t-test, one-tailed) at 95% (°) and 99% (•) against the original sd method.

		2007	2008
Run	Truncation	$F_1@K$	$F_1@K$
sd original	None	–	0.0681^{-}
B	Theoretical	0.0984	0.1361°
A	Technical	0.1011	0.1284°
highest		–	0.1848
median		–	0.0974
lowest		–	0.0051

5.2 Results and Discussion

We first discuss the overall quality of the rankings, and then the main topic of this paper—estimating the cut-off K.

The top half of Table 2 shows several measures on the two underlying rankings, **Legal07** and **Legal08**. We show precision at 5 (all top-5 results were judged by TREC); estimated recall at B; and the F_1 of the estimated precision and recall at R (i.e. the estimated number of relevant documents).

To determine the quality of our rankings in comparison to other systems, we show the highest, lowest, and median performance of all submissions in the bottom half of Table 2. As it turns out, **Legal08** obtains exactly the median performance for $Recall@B$ and $F_1@R$ when using all relevant documents in evaluation. Both rankings fare somewhat better than the median at $Prec@5$ and in evaluating with the highly relevant documents only. It is clear that our rankings are far from optimal in comparison with the other submissions. On the negative side, this limits the performance of the s-d method. On the plus side, it makes our rankings good representatives of the median-quality ranking.

Table 3 shows the results for the various thresholding methods. We see that the original s-d method stays well behind the $F_1@R$ in Table 2. Although this comparison is unfair, the mean estimated number of relevant items is generally not known, we expected the original s-d method to do better.

All runs with the improved version of the s-d method lead to significantly better results. The **B** run use the *theoretical*

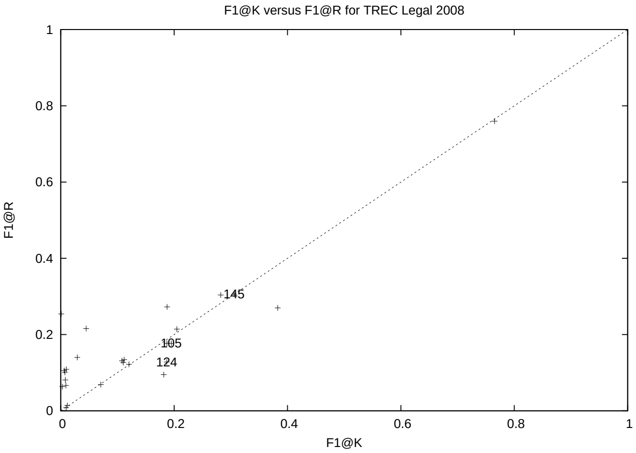


Figure 6: $F_1@R$ versus $F_1@K$ as estimated by s-d method for all 26 topics of TREC Legal 2008.

truncation of Section 3.3.1, whereas the **A** runs use the *technical truncation* of Section 3.3.2. For 2007, the technically truncated model **A** is superior to the theoretically truncated model **B**. For 2008, the technically truncated **A** model lags somewhat behind the theoretically truncated **B** model. In comparison with the 'old' non-truncated model, corresponding to our official TREC 2008 submission, both the truncated models obtain significantly better results.

We also show the highest, lowest, and median performance over the 23 submissions to TREC Legal 2008 (recall that the thresholding task is new at TREC 2008, so there is no comparable data for 2007). Note that the actual value of $F_1@K$ is a result of both the quality of the underlying ranking *and* choosing the right threshold. As seen earlier, our ranking has the median $Recall@B$ and $F_1@R$. With the estimated threshold of the s-d model, the $F_1@K$ is 0.1374, well above the median score of 0.0974.

There is still amble room for improvement. The $F_1@R$ in Table 2 is 0.1328 for 2007 and 0.1709 for 2008, and we obtain 75-80% of these scores. Obviously, R is not known in an operational system, and $F_1@R$ serves as a soft upper-bound on performance.

5.3 Further Analysis

Figure 6 show the F_1 scores of the Legal 2008 **B** run, plotted against the "ceiling" of F_1 at the estimated R. We will look in detail at some of the topics from 2007 and 2008 **B** runs:

Topic 73 B = 4,085; $est.R$ = 31,894; K_{opt} = 22,091.

Topic 105 B = 36,549; $est.R$ = 34,424; K_{opt} = 49,439.

Topic 124 B = 86,075; $est.R$ = 20,083; K_{opt} = 44,524.

Topic 145 B = 40,315; $est.R$ = 91,790; K_{opt} = 82,806.

Figure 7 compares the prediction of the s-d model with the official evaluation's estimated precision, recall, and F_1. Before discussing each of the topics in detail, an immediate observation is that the estimated (non-interpolated) precision is strikingly different from monotonically declining "ideal" precision curves.

For Topic 73 (Legal 2007), the estimated R exceeds the length of the ranking, and the K_{opt} corresponds to the last found relevant document at rank 22,091. The s-d model is clearly aiming too low and estimates R at 2,720 and K at 2,593.

Topic 105 (Legal 2008) has an R of 34,424, well within the length of the ranking, and the s-d model estimates an R of 36,503, near to the real R, and an estimated K of 28,952. The divergence in the prediction of K may be explained, in part, by the fact that K_{opt} always corresponds to a point where a relevant document is retrieved, and judged documents are very sparse down at this rank.

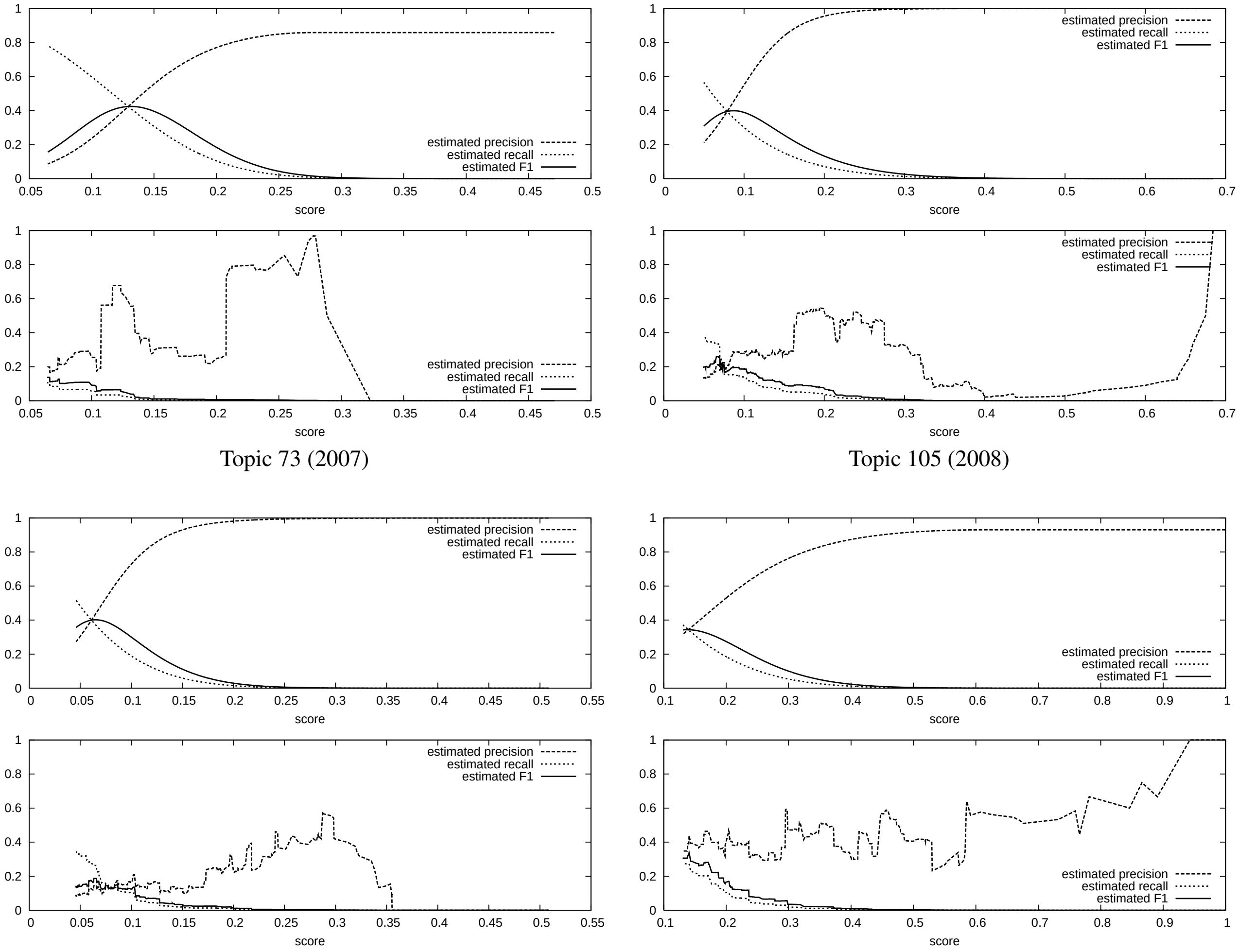
Topic 124 (Legal 2008) has an R of 20,083 and the s-d model predicts an R of 51,231 and a K of 43,597. Here, the R is overestimated but the K is very close to the K_{opt}. Topic 145 (Legal 2008) has an R of 91,790, very close to the length of the ranking. The s-d model predict an R of 87,060 and a K of 91,590, both relatively close to the official evaluation especially when bearing in mind that the K_{opt} is again at the last relevant document in the whole ranking.

6 Conclusions

We studied the problem of finding an "optimal" point to stop reading a ranked list, by selecting thresholds that optimize the F_1-measure. The approach taken employs the score-distributional threshold optimization (s-d), a non-parametric method proven effective for binary classification in earlier years. We made significant theoretical and computational improvements over the original method, and identified room for further improvements.

The method uses no other input than the document scores of a standard retrieval run, fit a mixture of (possibly truncated) normal and exponential distributions (normal for relevant, and exponential for non-relevant document scores), and calculate the optimal score threshold given the estimated distributions and their contributing weight. The experiments confirm that the s-d method is effective for determining thresholds, although there is still clear room for improvement: the effectiveness varies considerably per topic, with an average performance of 75-80% of $F_1@R$.

Assuming that a normal-exponential mixture is a good approximation for score distributions and that no relevance information is available, we believe that the improved methods described in this paper are *a)* as general as possible, *b)* they deal with most known theoretical anomalies and practical difficulties, and consequently, *c)* they bring us closer to the performance ceiling of s-d thresholding. If the effectiveness is deemed unsatisfactory, further improvements of s-d thresholding should come from using alternative mixtures or training data. Nevertheless, some other mixtures may be more difficult—or even impossible—to estimate.



Topic 73 (2007)

Topic 105 (2008)

Topic 124 (2008)

Topic 145 (2008)

Figure 7: S-D model predictions (top plots) versus the official evaluation (bottom plots).

Acknowledgments We thank Nir Nussbaum for producing the rankings and submitting the official runs. We also benefited from discussions with Stephen Robertson.

This research was supported by the Netherlands Organization for Scientific Research (NWO, grant # 612.066.513, 639.072.601, and 640.001.501).

References

[1] A. Arampatzis. Unbiased s-d threshold optimization, initial query degradation, decay, and incrementality, for adaptive document filtering. In *TREC*, 2001.

[2] A. Arampatzis and A. van Hameren. The score-distributional threshold optimization for adaptive binary classification tasks. In *Proceedings SIGIR'01*, pages 285–293, 2001.

[3] A. Arampatzis, J. Beney, C. H. A. Koster, and T. P. van der Weide. Incrementality, half-life, and threshold optimization for adaptive document filtering. In *TREC*, 2000.

[4] A. Arampatzis, J. Kamps, M. Koolen, and N. Nussbaum. Access to legal documents: Exact match, best match, and combinations. In *TREC*. NIST, 2007.

[5] A. Arampatzis, J. Kamps, and S. Robertson. Threshold optimization using truncated score distributions. *Unpublished*, 2009.

[6] C. Baumgarten. A probabilitstic solution to the selection and fusion problem in distributed information retrieval. In *Proceedings SIGIR '99*, pages 246–253. ACM Press, 1999.

[7] A. Bookstein. When the most "pertinent" document should not be retrieved – an analysis of the Swets model. *Information Processing and Management*, 13(6):377–383, 1977.

[8] K. Collins-Thompson, P. Ogilvie, Y. Zhang, and J. Callan. Information filtering, novelty detection, and named-page finding. In *TREC*, 2002.

[9] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[10] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*, volume 1. Wiley, 2nd edition, 1994.

[11] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*, volume 2. Wiley, 2nd edition, 1995.

[12] K. H. Knuth. Optimal data-based binning for histograms,

2006. URL http://arxiv.org/abs/physics/0605197v1.

[13] D. D. Lewis. Evaluating and optimizing autonomous text classification systems. In *Proceedings SIGIR'95*, pages 246–254. ACM Press, 1995.

[14] R. Manmatha, T. M. Rath, and F. Feng. Modeling score distributions for combining the outputs of search engines. In *Proceedings SIGIR'01*, pages 267–275, 2001.

[15] NIST/SEMATECH. e-handbook of statistical methods, 2008. http://www.itl.nist.gov/div898/handbook/.

[16] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 2nd edition, 1984.

[17] S. Robertson. On score distributions and relevance. In *Proceedings of 29th European Conference on IR Research, ECIR'07*, pages 40–51. Springer, Berlin, 2007.

[18] S. Robertson and J. Callan. Routing and filtering. In E. M. Voorhees and D. K. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*, chapter 5, pages 99–121. MIT Press, Cambridge MA, 2005.

[19] S. E. Robertson. The parametric description of retrieval tests. part 1: The basic parameters. *Journal of Documentation*, 25 (1):1–27, 1969.

[20] J. A. Swets. Information retrieval systems. *Science*, 141 (3577):245–250, 1963.

[21] J. A. Swets. Effectiveness of information retrieval methods. *American Documentation*, 20:72–89, 1969.

[22] Y. Zhang and J. Callan. Maximum likelihood estimation for filtering thresholds. In *Proceedings SIGIR'01*, pages 294–302. ACM Press, 2001.

A Chi-Square Goodness of Fit

To determine the quality of the fits, we bin the scores and calculate the χ^2 statistic

$$\chi^2 = \sum_i \frac{|O_i - E_i|^2}{E_i} \qquad (17)$$

where O_i and E_i are the observed and expected frequencies respectively for bin i. The expected frequency is calculated by

$$E_i = t\left(F(s_{i,a}) - F(s_{i,b})\right)$$

where $s_{i,a}$ and $s_{i,b}$ are respectively the lower and upper score limits of bin i, and $F(s) = (1-G_t)F(s|0) + G_tF(s|1)$ is the cumulative distribution function of the mixture under estimation.

The statistic follows, approximately, a χ^2 distribution with $M - 4 - 1$ degrees of freedom, where M is the number of bins and 4 is the number of parameters we estimate. The null hypothesis H_0 is that the observed data follow the estimated mixture. H_0 is rejected if the χ^2 of the fit is above the critical value of the corresponding χ^2 distribution at a significance level of 0.05 [15].

For the χ^2 approximation to be valid, E_i should be at least 5, thus we may combine bins in the right tail when $E_i < 5$. When the last E_i does not reach 5 even for $b = +\infty$, we only then apply the Yates' correction, i.e. subtract 0.5 from the absolute difference of the frequencies in Equation 17 before squaring.

Different fits on the same data can result to slightly different degrees of freedom due to combining bins. To compare the quality of different fits, so we can keep track of the best one irrespective its H_0 status, we use the χ^2 *upper-probability*; the higher the probability, the better the fit. As an initial upper-probability reference, we use the one of an exponential-only fit, produced by setting $\lambda = 1/(\mu_s - s_t)$.

The χ^2 statistic is sensitive to the choice of bins.

A.1 Score Binning

For binning, we use the optimal number of bins as this is given by the method described in [12]. The method considers the histogram to be a piecewise-constant model of the underlying probability density. Then, it computes the posterior probability of the number of bins for a given data set. This enables one to objectively select an optimal piecewise-constant model describing the density function from which the data were sampled. For practical reasons, we cap the number of bins to a maximum of 200.

B Formulas and Derivations

For completeness, we give here the rest of the formulas not given throughout the paper, and the derivations of those not found in the literature.

B.1 Density Functions

- standard normal distribution [16]:

$$\phi(s) = \frac{\exp\left(-s^2/2\right)}{\sqrt{2\pi}} \qquad s \in \mathbb{R} \qquad (18)$$

- exponential distribution [16]

$$\psi(s;\lambda) = \lambda \exp(-\lambda s) \qquad \lambda > 0,\ s \geq 0 \qquad (19)$$

B.2 Cumulative Distribution Functions

- standard normal [16]:

$$\Phi(s) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{s}{\sqrt{2}}\right)\right] \qquad s \in \mathbb{R} \qquad (20)$$

  where erf(.) is the *error function*.

- two-side truncated normal [10, pp.156–162]:

$$F(s|1) = \frac{\Phi\left(\frac{s-\mu}{\sigma}\right) - \Phi(\alpha)}{\Phi(\beta) - \Phi(\alpha)} \qquad s \in [s_{\min}, s_{\max}] \qquad (21)$$

  where α and β are given by Equation 11.

- exponential [16]:

$$\Psi(s;\lambda) = 1 - \exp(-\lambda s) \qquad s \geq 0 \qquad (22)$$

- shifted and right-truncated exponential:

$$F(s|0) = \frac{\Psi(s - s_{\min};\lambda)}{\Psi(s_{\max} - s_{\min};\lambda)} \qquad s \in [s_{\min}, s_{\max}] \qquad (23)$$

B.3 Moments of a Truncated Normal

These can be found in the literature, e.g. in [10]. Let S be a normally-distributed random variable with mean μ and variance σ^2, which we left-truncate at $s_{\min}$ and right-truncate at $s_{\max}$.

B.3.1 Expected Value

$$\mathrm{E}(S|s_{\min} \leq S < s_{\max}) = \mu + \frac{\phi(\alpha) - \phi(\beta)}{\Phi(\beta) - \Phi(\alpha)}\sigma \qquad (24)$$

We do not us the $\leq$ sign at the upper limit of S here (and in the equations below) to denote that the right-truncation is an option (i.e. $s_{\max}$ can be $+\infty$) in the context of this paper.

B.3.2 Variance

$$\mathrm{V}(S|s_{\min} \le S < s_{\max}) =$$

$$= \sigma^2 \left[1 + \frac{\alpha\,\phi(\alpha) - \beta\,\phi(\beta)}{\Phi(\beta) - \Phi(\alpha)} - \left(\frac{\phi(\alpha) - \phi(\beta)}{\Phi(\beta) - \Phi(\alpha)} \right)^2 \right] \tag{25}$$

B.4 Moments of a Shifted Truncated Exponential

We have not found those in the literature. Let S be an exponentially distributed random variable with rate parameter λ, which we shift by $s_{\min}$ and right-truncate at $s_{\max}$.

B.4.1 Expected Value

From the definition of the expected value of a truncated distribution[7] and Equation 19

$$\mathrm{E}(S|s_{\min} \le S < s_{\max}) = \frac{\int_{s_{\min}}^{s_{\max}} s\,\psi(s - s_{\min}; \lambda)\,\mathrm{d}s}{\Psi(s_{\max} - s_{\min}; \lambda)} =$$

$$= \frac{\lambda \exp(\lambda s_{\min})}{\Psi(s_{\max} - s_{\min}; \lambda)} \int\limits_{s_{\min}}^{s_{\max}} s \exp(-\lambda s)\,\mathrm{d}s$$

where the shift of the exponential by $s_{\min}$ is already taken into account. From lists of integrals of exponential functions[8]

$$\int\limits_{s_{\min}}^{s_{\max}} s \exp(-\lambda s)\,\mathrm{d}s = \left[\frac{\exp(-\lambda s)}{-\lambda} \left(s - \frac{1}{-\lambda} \right) \right]_{s_{\min}}^{s_{\max}}$$

Putting the last 2 equations together and working out the calculation leads to

$$\mathrm{E}(S|s_{\min} \le S < s_{\max}) = \frac{1}{\lambda} - \frac{s_{\max} \exp(-\lambda(s_{\max} - s_{\min})) - s_{\min}}{\Psi(s_{\max} - s_{\min}; \lambda)} \tag{26}$$

For only shift but no truncation ($s_{\min} \ne 0$, $s_{\max} = +\infty$), $\psi(s_{\max} - s_{\min}; \lambda) = 0$ and $\Psi(s_{\max} - s_{\min}; \lambda) = 1$, so Equation 26 becomes

$$\mathrm{E}(S|s_{\min} \le S) = \frac{1}{\lambda} + s_{\min}$$

which for a zero shift ($s_{\min} = 0$) it becomes $\mathrm{E}(S) = 1/\lambda$, as expected [16].

B.4.2 Variance

We can break down a shifted S to a mixture of its right-truncated and left-truncated parts weighted by a and b where $a + b = 1$. The two parts are non-correlated, so for their variances it holds that

$$\mathrm{V}(S|s_{\min} \le S) = a^2 \mathrm{V}(S|s_{\min} \le S < s_{\max}) + b^2 \mathrm{V}(S|s_{\max} \le S)$$

$$\Rightarrow \mathrm{V}(S|s_{\min} \le S < s_{\max}) = \frac{\mathrm{V}(s_{\min} \le S) - b^2 \mathrm{V}(S|s_{\max} \le S)}{a^2}$$

Since shifts do not affect variances, $\mathrm{V}(S|s_{\min} \le S) = \mathrm{V}(S|s_{\max} \le S) = 1/\lambda^2$. Moreover, $a = \Psi(s_{\max} - s_{\min})$, leading to

$$\mathrm{V}(S|s_{\min} \le S < s_{\max}) = \frac{1}{\lambda^2} \left(\frac{2}{1 - \exp(\lambda(s_{\min} - s_{\max}))} - 1 \right) \tag{27}$$

For only shift but no truncation ($s_{\min} \ne 0$, $s_{\max} = +\infty$), $\exp\left(\lambda\left(s_{\min} - s_{\max}\right)\right) = 0$ and Equation 27 becomes

$$\mathrm{V}(S|s_{\min} \le S) = \frac{1}{\lambda^2} = \mathrm{V}(S)$$

as expected; the shift does not affect the variance [16].

[7] http://en.wikipedia.org/wiki/Truncated_distribution

[8] http://en.wikipedia.org/wiki/List_of_integrals_of_exponential_functions