Where to Stop Reading a Ranked List?

Threshold Optimization using Truncated Score Distributions

Avi Arampatzis[1] Jaap Kamps[1] Stephen Robertson[2]

[1] University of Amsterdam, The Netherlands
[2] Microsoft Research Cambridge, United Kingdom

{avi,kamps}@science.uva.nl ser@microsoft.com

ABSTRACT

Ranked retrieval has a particular disadvantage in comparison with traditional Boolean retrieval: there is no clear cut-off point where to stop consulting results. This is a serious problem in some setups. We investigate and further develop methods to select the rank cut-off value which optimizes a given effectiveness measure. Assuming no other input than a system's output for a query—document scores and their distribution—the task is essentially a score-distributional threshold optimization problem. The recent trend in modeling score distributions is to use a normal-exponential mixture: normal for relevant, and exponential for non-relevant document scores. We discuss the two main theoretical problems with the current model, support incompatibility and non-convexity, and develop new models that address them. The main contributions of the paper are two truncated normal-exponential models, varying in the way the out-truncated score ranges are handled. We conduct a range of experiments using the TREC 2007 and 2008 Legal Track data, and show that the truncated models lead to significantly better results.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Filtering, and Retrieval Models

General Terms

Theory, Performance, Experimentation

1. INTRODUCTION

Ranked retrieval has a particular disadvantage in comparison to traditional Boolean retrieval: there is no clear cut-off point where to stop consulting results. This is hardly a practical problem in settings such as Web Search where only the initially retrieved results are consulted. In recall-oriented retrieval setups, such as searching patents or litigation and regulatory documents, the problem surfaces at full force. It is simply too expensive to give a ranked list with zillions of results to patent experts or litigation support professionals paid by the hour. This may be one of the reasons why ranked retrieval has been adopted very slowly in professional search.

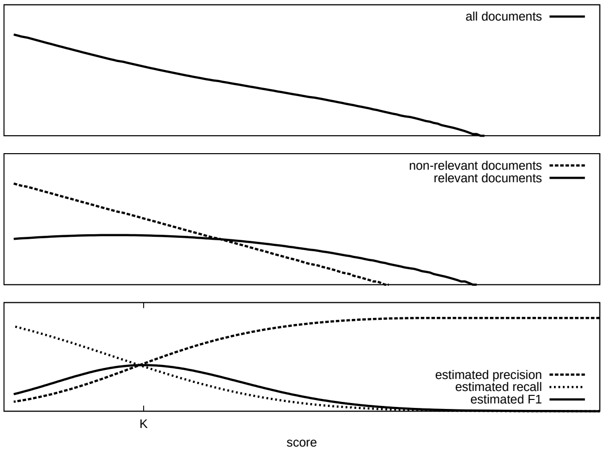

Figure 1: Score distribution (top); fitted mixture (middle); and the estimated Precision, Recall, and F_1 based on the fits (bottom).

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. At TREC 2008 the Legal Track addressed this problem head-on by requiring participants to submit the rank at which precision and recall are best balanced [12]. Our overall aim is to investigate and further develop methods to selecting a rank cut-off value K, per topic, for optimizing the given F_1-measure. Note that the resulting $F_1@K$ is as much as a result of the quality of the underlying ranking as of the choice of the cut-off value, but we focus entirely on determining the optimal rank cut-off values. The methods can be applied to a range of effectiveness measures; the measure under optimization is merely a parameter.

How can one select such a rank threshold given a ranked list? Selecting K is essentially equivalent to thresholding in binary classification or filtering [15]. Provided that there exist appropriate training data, a natural candidate line of approaches would be to apply machine learning [10, 17]. However, we assume here that results for a query are based solely on analysing the system output for this query—the scores and their distribution—and nothing else. In this context, machine learning methods cannot be applied and native IR methods (or "homegrown" methods [15]) must be used. Robertson and Callan [15] stress both the importance and the difficulty of threshold setting in homegrown IR methods. We opt for a pure IR approach: score-distributional threshold optimization (illustrated in Figure 1). In a nutshell, we treat the total score distribution as a weighted sum of the distributions of relevant and non-relevant scores. If we can recover these two distributions and their

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
Copyright 2009 ACM 978-1-60558-483-6/09/07 ...$5.00.

mix weight, we can approximate the numbers of relevant and non-relevant documents above and below any rank, and consequently, the total number of relevant documents R. Having these, any of the usual measures based on document counts can be calculated at all ranks, and the optimal rank for the selected measure can be found.

Apart from the practical use to determine a rank cut-off value optimizing some measure, the score-distributional methods also have theoretical importance. Provided that a model fits well enough, it can reveal the underlying relevance distribution (in much greater detail than through some judged documents), can transform scores to probabilities of relevance, and may reveal suboptimal behavior of the used ranking method. For example, these methods help us understand the underlying IR problem: What *is* the optimal cut-off point? How does it relate to R? Hence, advancing the score-distributional methods has direct relevance to IR theory. This is exactly why score distributions have attracted a lot of attention since the early days of IR. A range of known distributions have been considered, with the mixture of normal-exponential being the most popular in recent years. This mixture will be our starting point.

Over the years, two main problems have been identified with the normal-exponential mixture, and the main contribution of this paper is the development of new models that address these theoretical problems. The mixture, 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 minimum score, the normal has a full real axis support. This is a theoretical problem that we will address by investigating *truncated* distributional models.

- Can we develop truncated normal-exponential models?
  - Do these result in better goodness-of-fit?
  - Do these result in better thresholding?

From the point of view of how scores or rankings of IR systems should behave, Robertson [14] formulates the recall-fallout convexity hypothesis, a condition on how those two measures should trade-off for good systems. Similar conditions can be formulated on other quantities, e.g. on smoothed precision or probability of relevance which both should be monotonically increasing with the score. The normal-exponential mixture violates such conditions, thus it has a problem of *non-convexity*. One could seek alternative models without a convexity problem, but here our goals are more modest; we will at least address the convexity problem present in the mixture of normal-exponential distributions.

- Does the truncation remove some or all convexity problems?
  - How often does non-convexity occur in the observed score range?
  - What is the impact on the thresholding task?

The rest of this paper is organized as follows. In Section 2, we discuss earlier work on thresholding, in particular the score-distributional (s-d) method, and how it can be adapted for rank thresholding. The normal-exponential mixture is discussed in Section 3, and its theoretical problems in Section 4. Then, we develop new truncated mixture models in Section 5, and discuss the methods we use for their parameter estimation in Section 6. In Section 7, we conduct experiments on the TREC Legal Track data, both analyzing the resulting fits of the s-d models and the effectiveness of selecting rank cut-off values optimizing the F_1 measure used in the Legal 2008. Finally, we summarize the findings in Section 8.

2. S-D THRESHOLD OPTIMIZATION

In this section, we discuss earlier work on thresholding and score distributions, focusing on the *score-distributional threshold optimization*, a method first introduced at the TREC-9 Filtering Track [3, 4]. We re-formulate the method in order to stress variables related to the task we are dealing with, such as the total number of relevant documents R, as well as to clarify the assumptions under which the method works. Finally, we also adapt the method to the task of rank thresholding.

2.1 Score 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, 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 $> s$ 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 [15]. 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

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

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, R, and even normalizing scores to probabilities of relevance by straightforwardly applying the Bayes' rule [3, 11].

2.2 Rank Threshold 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 to simplify 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.

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:

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

Now 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. Score distributions have been modeled since the early years of IR with various known distributions. Swets [18] used two normal distributions, and later two exponentials [19]. Bookstein [6] used two Poisson distributions, and Baumgarten [5] used two Gamma distributions. Arampatzis et al. [4] started using a mixture of normal-exponential distributions: normal for relevant, exponential for non-relevant. Since this mixture has become the trend in the last few years [1, 3, 7, 11, 20], it is our starting point.

The normal-exponential model works as follows. Let us consider a general retrieval model which in theory produces scores in $[s_{\min}, s_{\max}]$, where $s_{\min} \in \mathbb{R} \cup \{-\infty\}$ and $s_{\max} \in \mathbb{R} \cup \{+\infty\}$. 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 $s_{\min} \in \mathbb{R}$. The two densities are given by

$$P(s|1) = \frac{1}{\sigma}\,\phi\left(\frac{s-\mu}{\sigma}\right) \qquad \sigma > 0,\ \mu, s \in \mathbb{R} \tag{6}$$

$$P(s|0) = \psi(s - s_{\min}; \lambda) \qquad \lambda > 0,\ s \geq s_{\min} \tag{7}$$

where $\phi(.)$ is the density function of the standard normal distribution, i.e. with a mean of 0 and standard deviation of 1, and $\psi(.)$ is the standard exponential density [13]:

$$\phi(s) = \frac{\exp\left(-s^2/2\right)}{\sqrt{2\pi}} \qquad s \in \mathbb{R} \tag{8}$$

$$\psi(s; \lambda) = \lambda \exp(-\lambda s) \qquad \lambda > 0,\ s \geq 0 \tag{9}$$

The corresponding cdfs are given by

$$F(s|1) = \Phi(s) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{s}{\sqrt{2}}\right)\right] \qquad s \in \mathbb{R} \tag{10}$$

$$F(s|0) = \Psi(s; \lambda) = 1 - \exp(-\lambda s) \qquad s \geq 0 \tag{11}$$

where $\operatorname{erf}(.)$ is the *error function* [13]. The total score distribution is written as

$$P(s) = (1 - G_n)\,P(s|0) + G_n P(s|1) \tag{12}$$

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

Despite its popularity, it was pointed out recently that the mixture of normal-exponential presents a theoretical anomaly in the context of IR. In practice, nevertheless, it has stood the test of time in the light of *a)* its (relative) ease to calculate, *b)* good experimental results, and *c)* 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 [3]. Also, long queries [3] or good queries/systems [11] seem to help.

In this paper, we do not set out to investigate alternative mixtures. We theoretically extend and refine the current normal-exponential model in order to address the problems which we will discuss in the next section.

4. NON-CONVEXITY

Over the years, two main problems of the normal-exponential model have been identified. Although we already generalized it 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 in the next section. In the remainder of this section, we will describe the other main problem: recall-fallout non-convexity.

From the perspective of how scores or rankings of IR systems should be, Robertson [14] formulates the recall-fallout convexity hypothesis:

> *For all good systems, the recall-fallout curve (as seen from the ideal point of 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 (which calculates precision only at points where relevant documents are found and interpolates in between). 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 [16]. 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 practical importance, the top of the ranking is very significant. The problem is a manifestation of the fact that an exponential tail extends further than a normal one. We focus on the problem at the top scores, and denote the lowest offending score with s_c. 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 observed 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 a better effectiveness than the random re-ordering. 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 be actually damaging to performance. If it does not, then the problem is inherent in the scoring formula producing

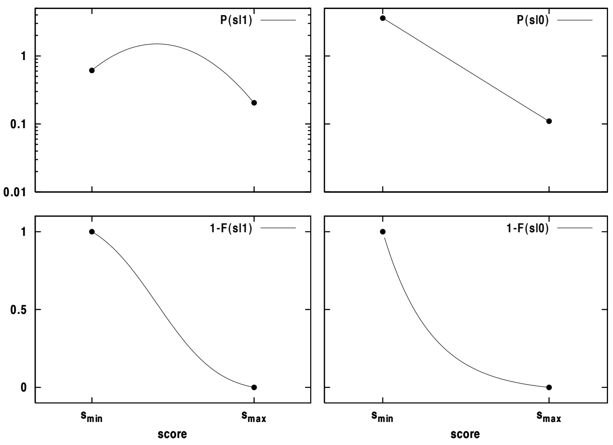

Figure 2: Theoretical truncation.

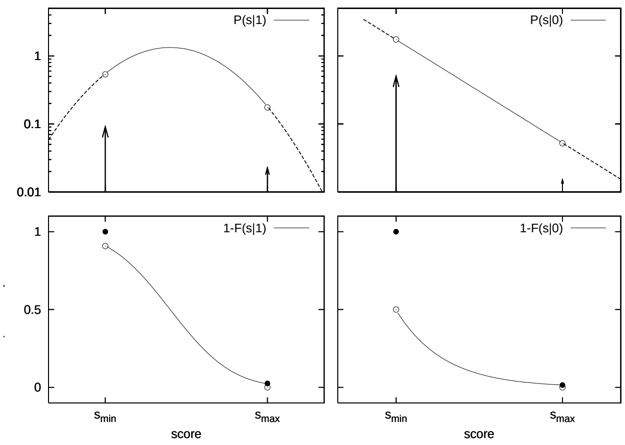

Figure 3: Technical truncation.

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 to determine whether a single ranking satisfies the PRP; precision for single queries is erratic, especially at early ranks, justifying the use of interpolated precision. According to interpolated precision all rankings satisfy the PRP, but that is due to the interpolation. 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. 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.

The models we introduce next, although they do not eliminate the problem, do not always violate such conditions imposed by the PRP (irrespective of whether it holds or not); a modest and conservative theoretical improvement over the original model which always does.

5. TRUNCATED S-D MODELS

In this section, we introduce two truncated normal-exponential models, with support restricted to $[s_{\min}, s_{\max}]$. The two models differ in the way the out-truncated score ranges are handled. In Section 6 we provide appropriate estimation methods for these models.

In order to enforce support compatibility, we introduce 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, 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, or by normalizing with the per-query score range. 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.

5.1 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:

$$P(s|1) = \frac{\frac{1}{\sigma}\,\phi\left(\frac{s-\mu}{\sigma}\right)}{\Phi(\beta) - \Phi(\alpha)} \qquad s \in [s_{\min}, s_{\max}] \tag{13}$$

$$P(s|0) = \frac{\psi(s - s_{\min}; \lambda)}{\Psi(s_{\max} - s_{\min}; \lambda)} \qquad s \in [s_{\min}, s_{\max}] \tag{14}$$

where

$$\alpha = \frac{s_{\min} - \mu}{\sigma} \qquad \beta = \frac{s_{\max} - \mu}{\sigma} \tag{15}$$

and $\phi(.)$, $\psi(.)$, $\Phi(.)$, $\Psi(.)$, are given by Equations 8–11. The cdfs of the above $P(s|1)$ and $P(s|0)$ are given by [9, 13, pp.156–162]:

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

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

5.2 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, calculating the ones at the left side requires to assume that the exponential extends below $s_{\min}$ to some new minimum score $s'_{\min}$:

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

The cdfs corresponding to the above densities are:

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

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

The equations in this section simplify somewhat when estimating their parameters from down-truncated ranked lists, as we will see in Section 6.1.

5.3 Relationship between the Models

For both models the right truncation is optional. For $s_{\max} = +\infty$, we get $\Phi(\beta) = \Psi(s_{\max} - s'_{\min}; \lambda) = 1$, leading to left-truncated models; this accommodates retrieval models with scoring support in $[s_{\min}, +\infty)$, $s_{\min} \in \mathbb{R}$. 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 an exponential.

When $s_{\min} \ll \mu \ll s_{\max}$ then $\Phi(\alpha) \approx 0$ and $\Phi(\beta) \approx 1$. If additionally $s'_{\min} = s_{\min}$, then $\Psi(s_{\min} - s'_{\min}; \lambda) = 0$ and $\Psi(s_{\max} - s'_{\min}; \lambda) \approx 1$. 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 result 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 [3].

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 $s_c \geq s_{\max}$, 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 theoretical nature. In practise, we should be interested in what happens when $s_c < s_1$. As we will later see in the experiment's section, 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.

6. PARAMETER ESTIMATION

In this section, we will develop parameter estimation methods for the truncated models. 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 [4, 7, 20]. 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 [11]. In order to deal with the biased training data in filtering, the EM method was also later adapted and applied for thresholding tasks [1].[1] Nevertheless, EM was found to be "messy" and sensitive to its initial parameter settings [1, 11].

6.1 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 by all previous research using the standard model, despite that it may affect greatly the estimation. Also, 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. Let us see how the formulas change.

Let us assume that the truncation score is s_t. For both truncated models, we need to estimate a two-side truncated normal at s_t and $s_{\max}$, and a shifted exponential by s_t right-truncated at $s_{\max}$, with $s_{\max}$ possibly be $+\infty$, from a set of top-t scores. The formulas that should be used are Equations 13 and 14 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 differ 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, 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, Equations 2 and 3 must be re-written as

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

while Equations 4 and 5 remain the same. $F(s|1)$ and $F(s|0)$ are now the cdfs either of Section 5.1 or 5.2, depending on the model.

For the choice of the technically truncated model, if there are any scores equal to $s_{\max}$ or $s_{\min}$ they should be removed from the dataset; these belong to the discontinuous legs of the densities given in Section 5.2. In this case, t should be decremented accordingly.[2]

6.2 Expectation Maximization

EM is an iterative procedure which converges locally [8]. Finding a global fit depends on the initial settings of the parameters.

6.2.1 Update Equations

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

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

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

with $P(j|s)$ 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 12 (for G_t instead of G_n).

[1] Another method for producing unbiased estimators in filtering can be found in [20], but it requires relevance judgements.

[2] 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 simplifing 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 our 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, the impact of the approximation is reduced.

We initialize the equations as it will be described in Section 6.2.3, 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.

6.2.2 *Correcting for Truncation*

If we use the truncated densities (Equations 13 and 14) in the above update equations, the μ_{new} and σ^2_{new} 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 correct to the right values using simple approximations.

Using Equation 21 in the Appendix, at the end of each iteration we correct the calculated λ_{new} as

$$\lambda_{\text{new}} \leftarrow \left(\frac{1}{\lambda_{\text{new}}} + s_t + \frac{s_{\max}\exp(-\lambda_{\text{old}}(s_{\max} - s_t)) - s_t}{\Psi(s_{\max} - s_t; \lambda_{\text{old}})} \right)^{-1} \tag{16}$$

using the λ_{old} from the previous iteration as an approximation. Similarly, based on Equations 19 and 20 in the Appendix, we correct the calculated μ_{new} and σ^2_{new} as

$$\mu_{\text{new}} \leftarrow \mu_{\text{new}} - \frac{\phi(\alpha') - \phi(\beta')}{\Phi(\beta') - \Phi(\alpha')}\sigma_{\text{old}} \tag{17}$$

$$\sigma^2_{\text{new}} \leftarrow \sigma^2_{\text{new}} \left[1 + \frac{\alpha'\,\phi(\alpha') - \beta'\,\phi(\beta')}{\Phi(\beta') - \Phi(\alpha')} - \left(\frac{\phi(\alpha') - \phi(\beta')}{\Phi(\beta') - \Phi(\alpha')} \right)^2 \right]^{-1} \tag{18}$$

where

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

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. 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. 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.

6.2.3 *Initialization and Number of Runs*

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. The randomized approach worked well because we initialize and run EM multiple (up to 100) times, and select the fit with the least χ^2 with the observed score data.

Although randomizing the parameters in their whole possible ranges works well, we used an improved initialization by randomizing into more probable narrower ranges motivated by IR considerations. This improves efficiency by reducing the number of EM iterations and runs. Furthermore, χ^2 values largely depend on how the observed scores are binned. Due to length limitations, we do not expand here on these issues but refer the reader to [2].

Table 1: Ranking quality for the Legal 2007 and 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

7. EXPERIMENTS

In this section, we apply the new models on the experimental setup based on the TREC 2007 and 2008 Legal Tracks. We discuss the underlying retrieval runs, and analyse the fits resulting from the old and new models. Then, we look at the effectiveness of the models in selecting a rank cut-off value K (per topic) for optimizing the given F_1-measure.[3] Note that the resulting $F_1@K$ is as much as a result of the quality of the underlying ranking as of the choice of the cut-off. Since our focus is the thresholding problem, we use an off-the-shelf retrieval system: the vector-space model of Apache's LUCENE.

7.1 Retrieval 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, and the generic SMART stoplist. The 2007 rankings are truncated at 25k results.

Legal08 Same run as above, using the `RequestText` as query. The 2008 rankings are truncated at 100k items.

We first discuss the overall quality of the rankings. The top half of Table 1 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 (i.e. the size of the returned set of the reference Boolean run); 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 2008 submissions in the bottom half of Table 1.[4] As it turns out, **Legal08** obtains exactly the median performance for $Recall@B$ and $F_1@R$, and fares somewhat better than the median at $Prec@5$. 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 methods. On the positive side, our **Legal08** ranking is a good representative of the participating systems.

7.2 Convexity of Fits

We fit the two new models, as well as the old model, by estimating the appropriate parameters using EM as discussed above. Table 2 provides some data on the convexity of the resulting fits. We look at the number of topics where the fit (as measured by the χ^2 with the observed score data) improves over the non-truncated approach, and see that the fit improves for 80% of the topics. We also investigate the number of fits presenting the non-convexity anomaly within the observed score range, i.e. at a rank below rank 1 ($k_c > 1$). We see that the anomaly shows up in a large number of topics; 53-66% for the truncated models, but in almost all topics,

[3] More information about the collection, topics, and evaluation measures can be found in [12].

[4] We include these for 2008 to be able to compare to the thresholding task later (for which is there is no comparable data from 2007).

Table 2: The effects of truncation on the convexity of fits.

Year	Truncation	Improved fit	$k_c > 1$	$\widetilde{k_c}$	$k_c > \widetilde{R}$
2007	none	–	46 (92%)	47	3 (6%)
2007	theoretical	40 (80%)	33 (66%)	26	3 (6%)
2007	technical	40 (80%)	33 (66%)	29	5 (10%)
2008	none	–	43 (95%)	312	1 (2%)
2008	theoretical	37 (82%)	24 (53%)	563	0 (0%)
2008	technical	35 (78%)	29 (64%)	89	0 (0%)

92-95%, for the original non-truncated model. Thus, beyond the theoretical improvement of the truncated models not always violating convexity, truncation also helps in practice during parameter estimation resulting in a higher fraction of convex fits.

What is the impact of non-convexity on thresholding? By randomizing the affected ranks rather than re-ranking, the net effect is that the s-d method turns "blind" at rank numbers $< k_c$ restricting the estimated optimal thresholds with $K \geq k_c$. 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 for 2007 and 32,233 for 2008). Since $K < k_c$ is unlikely, thresholding quality should not be affected—on average. Nevertheless, for a small number of topics (2-10%), the problem appears for $k_c > \widetilde{R}$ and non-convexity should have a more significant impact. For a good fraction of such topics, a large k_c indicates a fitting problem rather than a theoretical one. Figure 4 illustrates this: the resulting fit (top plot) looks good but has a convexity problem in the whole ranking ($k_c \geq 25,000$), indicated by having to flatten its estimated precision in the whole range. The fit regards all highest scoring documents as non-relevant, and it could have been rejected on IR grounds: for example, by requiring the expected relevant score to be larger than the expected non-relevant. The next best fit (bottom plot) has no convexity problem. Overall, our data suggest that non-convexity has an insignificant impact on s-d thresholding.

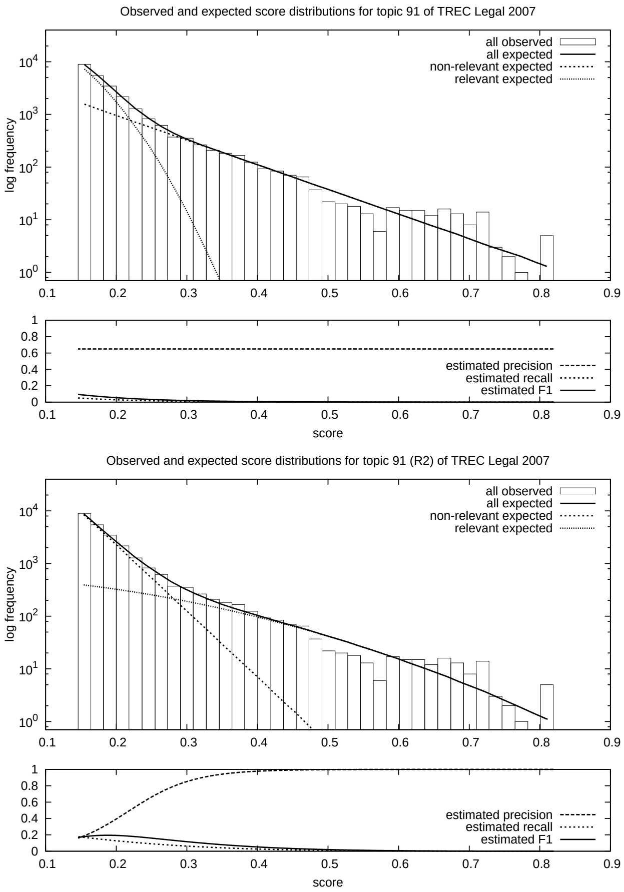

Figure 4: For topic 91 (top plot), the fit looks good but has a convexity problem in the whole ranking. The second best fit (bottom plot) has no convexity problem.

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.

Truncation	$F_1@K$	
	2007	2008
none	0.0751	0.0744
theoretical	0.1069°	$0.1356^{\bullet}$
technical	0.1032°	$0.1362^{\bullet}$
highest	–	0.1848
median	–	0.0974
lowest	–	0.0051

7.3 Thresholding

For the threshold optimization we simply use the fitted mixture of normal and exponential, and calculate the rank that maximizes the F_1 measure. Note that a 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. We have three runs corresponding to the use of truncation:

none Runs using the original non-truncated s-d model.

theoretical Runs using the *theoretical truncation* of Section 5.1.

technical Runs using the *technical truncation* of Section 5.2.

Table 3 shows the results for the various thresholding methods. All runs with the truncated s-d models lead to significantly better results than the old s-d model. For 2007, the theoretically truncated model scores better than the technically truncated model. For 2008, the technically truncated model gets a somewhat better score than the theoretically truncated model. Hence, it is not clear which of the truncation models is superior—the differences are not significant.

We also show the highest, lowest, and median performance over the 23 submissions to TREC Legal 2008 (the thresholding task is new at TREC 2008, so there are no comparable data for 2007). Note again 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 thresholds of the s-d model, the $F_1@K$ is 0.136, well above the median of 0.0974. There is still room for improvement. Although this comparison is unrealistic—the mean estimated number of relevant items is generally not known—we achieve up to 80% of the $F_1@R$ of Table 1.

8. DISCUSSION AND CONCLUSIONS

We studied the problem of finding an optimal point to stop reading a ranked list, by selecting thresholds that optimize a given measure. Assuming no other input than a system's output for a query—document scores and their distribution—the task is essentially a score-distributional threshold optimization problem. The recent trend in modeling score distributions is to use a normal-exponential mixture: normal for relevant, and exponential for non-relevant document scores. We discussed the two main theoretical problems with the current model—support incompatibility and non-convexity—and developed new models that address them.

The main contributions of the paper are two truncated normal-exponential models, varying in the way the out-truncated score ranges are handled. The theoretical truncation assumes that no data exist outside the truncated score range; the technical truncation assumes that the 'missing' data are accumulated at the two truncation points. We showed that truncation improves the goodness-of-fit for most topics and reduces non-convexity problems at the top of rankings, although the problem remains for a considerable fraction of topics. Our analysis revealed that some of the extreme cases can be attributed to fitting problems rather than problems of the underlying ranking or with the normal-exponential mixture, and suggested that such fits can be rejected on IR grounds (e.g. by requiring that the expected relevant score is larger than the expected non-relevant score). We also showed that for the overwhelming majority of the remaining topics non-convexity occurs at early ranks, where it has an insignificant impact on the s-d thresholding given the large numbers of relevant documents in the setup. This is confirmed in a range of experiments using the TREC 2007 and 2008 Legal Track data, where we showed that the truncated models lead to significantly better performance over the standard model.

Assuming that the 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 *a)* are as general as possible, *b)* deal with most known theoretical anomalies and practical difficulties, and consequently, *c)* bring us closer to the performance ceiling of s-d thresholding. Further improvements of s-d thresholding should come from using training data or alternative mixtures. Although we focused on the normal-exponential mixture, truncated models can also be defined for other distributions. Since all retrieval runs tend to be truncated for practical reasons, truncation is an important factor for fitting any distribution.

9. REFERENCES

[1] A. Arampatzis. Unbiased s-d threshold optimization, initial query degradation, decay, and incrementality, for adaptive document filtering. In *Proceedings TREC 2001*. NIST, 2002.

[2] A. Arampatzis and J. Kamps. Where to stop reading a ranked list? In *Proceedings TREC 2008*. NIST, 2009.

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

[4] 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 *Proceedings TREC 2000*. NIST, 2001.

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

[6] 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.

[7] K. Collins-Thompson, P. Ogilvie, Y. Zhang, and J. Callan. Information filtering, novelty detection, and named-page finding. In *Proceedings TREC 2002*. NIST, 2003.

[8] 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.

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

[10] D. D. Lewis. Applying support vector machines to the TREC-2001 batch filtering and routing tasks. In *Proceedings TREC 2001*, pages 286–292. NIST, 2002.

[11] 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.

[12] D. W. Oard, B. Hedin, S. Tomlinson, and J. R. Baron. Overview of the TREC legal track. In *Proceedings TREC 2008*. NIST, 2009.

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

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

[15] 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, 2005.

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

[17] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34:1–47, 2002.

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

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

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

Appendix

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

The moments of a truncated normal can be found in the literature [9]. Let S be a normally-distributed random variable with mean μ and variance σ^2, left-truncated at $s_{\min}$ and right-truncated at $s_{\max}$. Its expected value is

$$\mathrm{E}(S|s_{\min} \leq S < s_{\max}) = \mu + \frac{\phi(\alpha) - \phi(\beta)}{\Phi(\beta) - \Phi(\alpha)}\sigma \qquad (19)$$

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. For the variance:

$$\begin{aligned} &\mathrm{V}(S|s_{\min} \leq S < s_{\max}) \\ &\quad = \sigma^2 \left[1 + \frac{\alpha\,\phi(\alpha) - \beta\,\phi(\beta)}{\Phi(\beta) - \Phi(\alpha)} - \left(\frac{\phi(\alpha) - \phi(\beta)}{\Phi(\beta) - \Phi(\alpha)} \right)^2 \right] \end{aligned} \qquad (20)$$

Concerning the expectation of a shifted truncated exponential, we have not found the formula 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 $s_{\max}$. From the definition of the expected value of a truncated distribution and Equation 9:

$$\begin{aligned} \mathrm{E}(S|s_{\min} \leq S < s_{\max}) &= \frac{\int_{s_{\min}}^{s_{\max}} s\,\psi(s - s_{\min}; \lambda)\,\mathrm{d}s}{\Psi(s_{\max} - s_{\min}; \lambda)} \\ &= \frac{\lambda \exp(\lambda s_{\min})}{\Psi(s_{\max} - s_{\min}; \lambda)} \int_{s_{\min}}^{s_{\max}} s \exp(-\lambda s)\,\mathrm{d}s \end{aligned}$$

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

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

Putting these equations together and working out the calculation:

$$\mathrm{E}(S|s_{\min} \leq S < s_{\max}) = \frac{1}{\lambda} - \frac{s_{\max} \exp(-\lambda(s_{\max} - s_{\min})) - s_{\min}}{\Psi(s_{\max} - s_{\min}; \lambda)} \qquad (21)$$

For only shift but no truncation (i.e. $s_{\min} \neq 0$ and $s_{\max} = +\infty$), $\psi(s_{\max} - s_{\min}; \lambda) = 0$ and $\Psi(s_{\max} - s_{\min}; \lambda) = 1$. Equation 21 becomes

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

which without a shift ($s_{\min} = 0$) becomes $\mathrm{E}(S) = 1/\lambda$, as expected [13].