# AVERAGE PRECISION AND THE PROBLEM OF GENERALISATION

**Simon I. Hill***, **Hugo Zaragoza**†, **Ralf Herbrich**†, **Peter J. W. Rayner***

**Department of Engineering*
*University of Cambridge*
*Cambridge, UK.*
`sih22,pjwr@cam.ac.uk`

*†Microsoft Research Ltd.*
*Cambridge, UK.*
`hugoz,rherb@microsoft.com`

Abstract

In this paper we study the problem of generalisation in information retrieval. In particular we study precision-recall curves and the average precision value. We provide two types of bounds: large-deviation bounds of the average precision and maximum deviation bounds with respect to a given point of the precision recall curve. The first type of bounds are useful to answer the question: how far can true average precision be from the value observed on a test collection? The second is useful for obtaining bounds on average precision when tight bounds on a particular point of the curve can be established, as is the case when training SVMs or Perceptrons for document categorisation.

Keywords

Information Retrieval, Precision, Recall.

## 1. INTRODUCTION

Information retrieval (IR) relies heavily on performance evaluation techniques. Numerous papers have been written comparing systems across a wide range of performance measures and corpora. Formally, however, we know very little about performance measures: how to they relate to each another and to characteristics of the corpora. More importantly, we know very little about the dependences between these measures and the different IR algorithms. The study of generalisation in statistics offers a framework for these problems. In this paper we are going to present a statistical analysis of the most commonly used performance measure in IR: the *precision-recall curve*. We will provide some answers to the following question: How far can *true* performance be from the performance *observed* on a particular test corpus?

There is a good reason why we did not need to worry about this problem in the past: Broadly speaking, simple systems generalise well. That is, any reasonable performance measure will yield consistent results across simple IR algorithms, because for these algorithms true performance cannot (provably) be too far from the performance observed in a test corpus [5]. However, as IR systems increase in complexity (i.e. dimensionality of feature space and number of parameters) the true performance of a system can fluctuate further away from its observed value.

This problem is well known in the fields of applied statistics, probabilistic inference and machine learning. In the past twenty years these fields have provided a wide range of concepts and techniques tackling the problem of generalisation. However, so far these studies have concentrated on the problem of *classification* and have rarely dealt with the problem of most interest in IR: *ranking*. In this paper we apply some statistical techniques to the problem of *document ranking* in order to study the precision-recall curve.

### 1.1 Precision and Recall

Let us describe the problem more formally. Consider a labelled collection $C := \{(x_i, y_i)\}_{i=1}^m$ where $x_i \in X$ is a document, $y_i \in \{-1, +1\} =: Y$ is a label ($+1$ meaning the document is relevant), $m$ is the size of the collection, $X$ is the set of all possible documents, and $m^+$ is the number of relevant documents in $C$.

We will assume that our collection $C$ is a sample of a much larger collection of labelled documents. In particular, we will consider that there is some probability distribution $\pi_{XY}$ defined over $X \times Y$ where the conditional distribution $\pi_{Y|X=x}$ is a deterministic function of the documents $x$ for a given query or topic.

Let us start by considering an IR system $h : X \to Y$ which, given a document $x$, outputs only one of two relevancy values $+1$ or $-1$. We can evaluate the performance of this system on our collection $C$ by counting the number of relevant documents correctly labelled (denoted by $m^{++}$) and the documents incorrectly labelled as relevant ($m^{+-}$). Then *precision*, $q_{C,h}$, and *recall*, $r_{C,h}$, are defined as as follows:

$$r_{C,h} := \frac{m^{++}}{m^+}, \tag{1}$$

$$q_{C,h} := \frac{m^{++}}{m^{++} + m^{+-}}. \tag{2}$$

Taking into account that our collection $C$ is a sample from $\pi_{XY}$ on can interpret these two measures as estimates, namely; $r_{C,h} \approx r_h := P(h(x) = +1 | y = +1)$, and $q_{C,h} \approx q_h := P(y = +1 | h(x) = +1)$.

Note that recall is identical to the so called *true positive* acceptance rate, or sensitivity. Precision however is quite an unusual measure: it is *not* the usual *false positive* acceptance rate or *specificity*, $s_h := P(h(x) = +1 | y = -1)$, traditionally used in ROC and OC curves. It can be shown that these two quantities are not linearly related. Furthermore, the *probability of misclassification error*, $P(h(x) \neq y)$, is non-linearly related to precision and recall as expressed in the following equation:

$$q_h = \left(1 - \frac{P(y=-1)}{P(y=+1)} \frac{1-r_h}{s_h}\right)^{-1}$$

## 1.2 Precision-Recall curves

In general, ranking functions are not binary but implement a real-valued function $f : X \to \mathbb{R}$ to rank documents with respect to their likelihood of relevancy. Given a collection $C$ and a ranking function $f$, we introduce the following shorthand notation:

- $(i) \in \{1, \ldots, m\}$ returns the index of the $i$th ranked document.
- $(j)^+ \in \{1, \ldots, m^+\}$ returns the index of the $j$th ranked document when only relevant documents are considered.

Let $f_j^+ := f\left(x_{(j)^+}\right)$ and $i_{(j)^+}$ the rank of $x_{(j)^+}$. In order to enhance the understanding consider the collection $C = \{(x_1, +1), (x_2, +1), (x_3, -1), (x_4, -1)\}$ and a ranking function $f$ such that $f(x_3) > f(x_2) > f(x_1) > f(x_4)$. Then $\left[x_{(1)} x_{(2)} x_{(3)} x_{(4)}\right] = [x_3, x_2, x_1, x_4]$, $\left[x_{(1)^+}, x_{(2)^+}\right] = [x_2, x_1]$, $\left[f_1^+, f_2^+\right] = [f(x_2), f(x_1)]$, and $\left[i_{(1)^+}, i_{(2)^+}\right] = [2, 3]$.

In order to evaluate the quality of the ranking produced by $f$ in $C$, we are going to use the previously defined measures of precision and recall. For this we introduce a threshold $b \in \mathbb{R}$ and construct the new classification function:

$$h_b(x) = \operatorname{sign}(f(x) - b) \ . \tag{3}$$

Precision and recall now depend on the value of $b$; There are $m$ different $(q_b, r_b)$ values[1], which can be obtained by the $m$ classifiers $\left\{h_{f(x_1)}, \ldots, h_{f(x_m)}\right\}$.

Consider the subset of these classifiers obtained by considering only the relevant documents, sorted with respect to $f$. Their recall values are:

$$\mathbf{r}(C, f) := \left[\frac{1}{m^+}, \frac{2}{m^+}, \ldots, \frac{m^+}{m^+}\right]$$

Plotting the corresponding precision values at these points we obtain the *precision recall curve*

$$(\mathbf{q}(C, f), \mathbf{r}(C, f)) := \left[\left(\frac{j}{i_{(j)^+}}, \frac{j}{m^+}\right)\right]_{j=1}^{m^+} ,$$

[1]With a slight abuse of notation, $q_b$ denotes $q_{C,h_b}$. The document collection $C$ should be clear from the context.

which gives us an indication of the performance of our system over all possible recall values. Finally, we can state the performance measure of interest, the *average precision,* which is the average point in the precision-recall curve:

$$A_f(C) := \frac{1}{m^+}\sum_{j=1}^{m^+} q_j(C,f) = \frac{1}{m^+}\sum_{j=1}^{m^+}\frac{j}{i_{(j)^+}}\,.$$

### 1.3 The problem of generalisation

We can now state the problem we are interested in more formally: We are trying to characterise the difference of the average precision over the random draw of two test collections $C$ and $C'$ (with high probability). Hence, we are interested in

$$P_{C,C'\sim\pi_{XY}^m}\left(\left|A_f(C)-A_f(C')\right|>\epsilon\right)\,.$$

Of special importance will be the quantity $A_f := \mathbb{E}_{C\sim\pi_{XY}^m}\left\{A_f(C)\right\}$, that is, the expected average precision over the random draw of document collections $C$. As we do not wish to characterise the sampling distribution $\pi_{XY}$, we will consider expectation over all subsets of documents of size $m$ and exactly $m^+$ positive points (as it will be seen later, this does not restrict our results in practice). For this reason, we will be interested in the quantity $A_f(m,k) := \mathbb{E}_{C\sim\pi_{XY}^m}\left\{A_f(C)\,|m^+=k\right\}$. As a consequence, our ultimate interest is in upper bounding

$$P_{C\sim\pi_{XY}^m}\left(A_f(C)-A_f(m,k)>\epsilon\;\middle|\;m^+=k\right)\,. \tag{4}$$

In other words: How far can the observed average precision $A_f(C)$ be from the expected average $A_f(m,m^+)$ for *any* function $f$ and any collection $C$?

## 2. BOUNDING AVERAGE-PRECISION

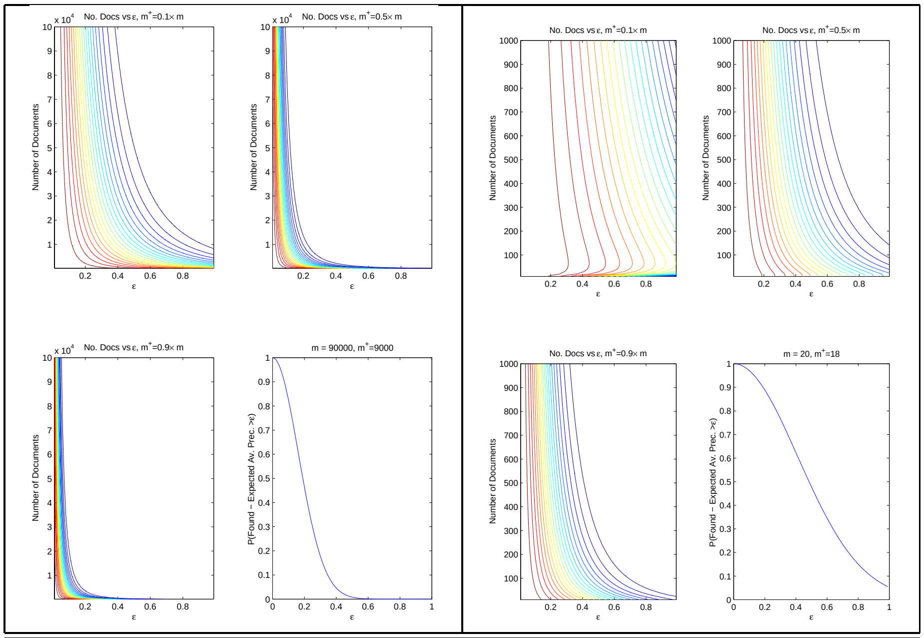
In order to answer the above question we will use McDiarmid's inequality, which allows us to bound the deviation of any function of a sample based on the maximum deviation that can be observed by this function when a single object of the sample is altered. Obviously, we would like to be able to show that these two quantities to grow closer as $m$ and $m^+$ grow larger. We will find probabilistic bounds on the difference between these two quantities and prove asymptotic convergence. These constitute the first average precision generalisation bounds that we are aware of.

In our notation, the results of McDiarmid reads as follows (see [2] or [4] for more details): For any $i$, let $\tau_i$ be defined such that

$$\forall(x,y):\forall C:m^+=k:\quad\left|A_f(C)\;-A_f\left(C_{i\leftrightarrow(x,y)}\right)\right|\le\tau_i\,, \tag{5}$$

where $C_{i\leftrightarrow(x,y)}$ is the document collection with the $i$th example replaced by $(x,y)$. In other words, $\tau_i$ bounds the maximal change in average precision for a ranking function $f$ if the $i$th document-class pair is replaced. Then, for all $\varepsilon>0$,

$$P_{C\sim\pi_{XY}^m}\left(A_f(C)-A_f(m,k)>\epsilon\;\middle|\;m^+=k\right)<\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^m\tau_i^2}\right)\,. \tag{6}$$




| $m \in \{1, \ldots, 100\,000\}$ | $m \in \{1, \ldots, 1\,000\}$ |
|---|---|

*Figure - 1*: Contour plot of (6) for $A_f(C)$. The probability that the expected average precision lies within $\epsilon$ of the found average precision is represented for a range of $\epsilon$, $m$ and six different cases of $m^+$.

## 2.1 The precision - recall curve

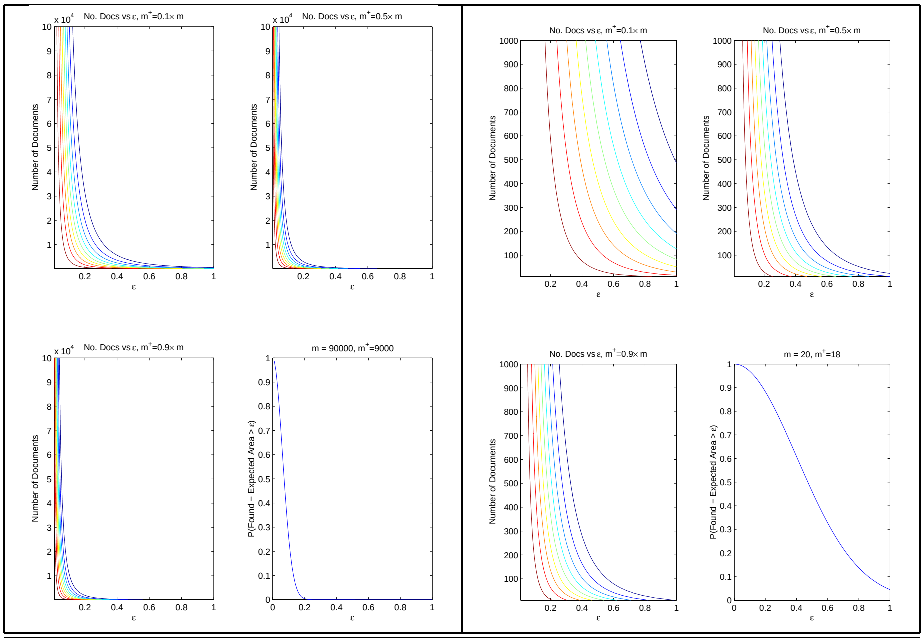
The derivation of $\tau_i$ is quite a laborious process for the case of the precision-recall curve and involves the study of several different cases; an overview is presented in Appendix A, section A.1. We show that

$$\tau_i \leq \frac{1}{m^+ + 1} \sum_{l=1}^{m^+} \frac{1}{l} \leq \frac{\ln(m^+) + B_E + \frac{1}{2m^+}}{m^+ + 1} \tag{7}$$

where $B_E$ denotes Euler's constant ($\approx 0.5772$, for more on this second bound see [1, p.3]). Hence (6) becomes,

$$P_{C \sim \pi_{XY}^m}\left(A_f(C) - A_f(m,k) > \epsilon \;\middle|\; m^+ = k\right) < \exp\left[-\frac{2\epsilon^2}{m}\left(\frac{\ln(m^+) + B_E + \frac{1}{2m^+}}{m^+ + 1}\right)^{-2}\right]$$

This demonstrates that the average precision observed on a test collection converges to the expected average precision as $m^+$ increases because $\tau_i \to 0$. More importantly, it can be shown that, if $\frac{m^+}{m}$ is constant, the observed average precision converges to the expected (or true) one as $m$ goes to infinity. This conclusion is supported by experiments, as shown in Figure *1*.




| $m \in \{1, \ldots, 100\,000\}$ | $m \in \{1, \ldots, 1\,000\}$ |
|---|---|

*Figure - 2*: Contour plot of (6) for $A'_f(C)$. The probability that the expected average precision lies within $\epsilon$ of the found average precision is represented for a range of $\epsilon$, $m$ and six different cases of $m^+$.

Figure *1* clearly illustrates convergence of the bound with both $m$ and $m^+$ for $m^+ \geq 4$. For extremely small $m^+$ it is clear from the expression that the limiting trends will be less influential than initial transients. Outside this region however the bound can be seen to consistently become tighter. This is demonstrated by the given example at $m = 90\,000$ and $m^+ = 9000$ which shows a confidence of more than 95% that $\epsilon \leq 0.4$.

## 2.2 An alternative to average precision

We propose here a modified average precision measure which is closely related to average precision, yet it is tightly bound by McDiarmid's inequality. With this we do not wish to replace average precision in practice, but rather to find an alternative measure that we can bound and under which we can study the effect of increasing sample size and varying ratio of relevant documents. This measure offers the advantage of using the usual precision measure (unlike OC curves), and empirically correlates strongly with average precision.

The alternative curve that we consider plots precision against rank number (as opposed to recall) for all ranked documents, not only the relevant documents. The new average precision, $A'_f(C)$, is given by:

$$A'_f(C) := \frac{\frac{1}{m}\sum_{i=1}^{m} \frac{m_i^{++}}{i}}{A^*} \tag{8}$$

where $m_i^{++}$ is the value of $m^{++}$ for $h_{f(x_{(i)})}$ (see (3)) and $A^* \frac{m^+}{m} \left(1 + \sum_{i=m^++1}^{m} \frac{1}{i}\right)$ is a normalising factor (the maximum possible value of the numerator). One key advantage of the revised curve is a decrease in volatility. For example, if the highest ranked document is irrelevant then it has have a profound effect on $A_f(C)$ but not on $A'_f(C)$.

The use of the curve is simply to obtain insight into the quality of the ranking and, in that sense, its choice is somewhat *ad hominem*. It may well be that such volatility is not an accurate representation of the true perception. This being the case the revised curve is proposed as a more stable alternative for study.

The same procedure used above to bound $A_f(C)$ can be used for $A'_f(C)$. In this case eight potential maximum average precision changes were found and these are listed in Appendix A, section A.3. Of these, using an approach similar to that in section A.2, *exceptis excipiendis*, the bound is given by

$$\tau_i \leq \frac{A^* \sum_{l=1}^{m} \frac{1}{l} - \frac{1}{mm^+}\left(\sum_{l=2}^{m^+} \frac{l}{l+m^-} + 2\right)}{A^* m \left(A^* - \frac{1}{mm^+}\right)} \tag{9}$$

and so, in this case, (6) becomes,

$$\begin{aligned} &P_{C\sim\pi_{XY}^m}\left(A'_f(C) - A'_f(m,k) > \varepsilon \;\middle|\; m^+ = k\right) \\ &\qquad < \exp\left[-\frac{2\varepsilon^2}{m}\left(\frac{A^* m \left(A^* - \frac{1}{mm^+}\right)}{A^* \sum_{l=1}^{m} \frac{1}{l} - \frac{1}{mm^+}\left(\sum_{l=2}^{m^+} \frac{l}{l+m^-} + 2\right)}\right)^2\right]. \end{aligned} \tag{10}$$

In Figure *2* we have numerically evaluated the bounds for different values of $m$ and $m^+$. We see that this new measure also converges but offers much tighter bounds. For example, for $m = 90\,000$ and 10% of relevant documents ($m^+/m = 0.1$), with probability at least 95%, we have that the observed average precision can be at most 20% larger than the true average precision.

It is clear that a very large number of positive documents is required before anything approaching a tight statistical bound can be obtained. This is to be expected, given that we have not used any information on the nature of the ranking function $f$. Nevertheless these bounds offer a first insight into the nature of average precision.

## 3. FROM POINT-BOUNDS TO CURVE-BOUNDS

Learning theory allows us to bound the misclassification error much more tightly (at least in principle) using quantities specific to our trained classifier $h$, such as its margin, its leave-one-out error or the fraction of documents used for training (see, e.g. [2]). In [3] leave-one-out type bounds are established on precision, recall and on the F1 measure. Unfortunately, these bounds apply to a single classifier $h_b$ and do not tell us how classifier $h_{b'}$, $b' \neq b$, will behave. This is of great importance because, in order to bound $A_f(C)$ we need to compute precision and recall bounds for *all* the $m^+$ classifiers $\{h_{f(x_1)}, \ldots, h_{f(x_m)}\}$. We know that one and only one of these classifiers will correspond to our particular classifier $h_b$, and therefore we can only bound a particular point of the precision recall curve.

In order to be able to use [3]'s result we study the best and worst precision-recall curves going through a particular point given by $m^{++}$, $m^{+-}$ (see (1) and (2)). Here, best and worst refer to maximum and minimum average precision, for a constant number $m$ of documents and $m^{+} = k$ relevant documents.

It can be proven by contradiction that the best (worst) precision-recall curve is obtained by ranking all relevant documents first (last). Following this argument, the maximum and minimum possible average precision are therefore given by:

$$\begin{aligned} A_f^{\mathrm{MAX}}(C) &:= \frac{1}{k}\left(m^{++} + \sum_{j=m^{++}+1}^{k} \frac{j}{m^{+-}+j}\right), \\ A_f^{\mathrm{MIN}}(C) &:= \frac{1}{k}\left(\sum_{j=1}^{m^{++}} \frac{j}{m^{+-}+j} + \sum_{j=m^{++}+1}^{k} \frac{j}{(m-k)+j}\right). \end{aligned}$$

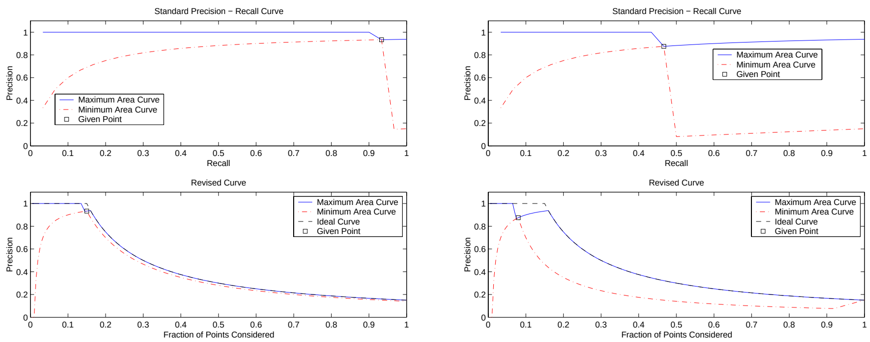
Similarly, for the alternative average precision measure, $A'_f(C)$,

$$\begin{aligned} A_f^{\prime,\mathrm{MAX}}(C) := \frac{1}{m\cdot A^*}&\left[m^{++}\left(1 + \sum_{i=m^{++}+1}^{m^{++}+m^{+-}} \frac{1}{i}\right)\right. \\ &\left. + \sum_{i=m^{++}+1}^{k} \frac{i}{i+m^{+-}} + \sum_{i=m^{--}+1}^{m-k} \frac{k}{i+k}\right], \\ A_f^{\prime,\mathrm{MIN}}(C) := \frac{1}{m\cdot A^*}&\left[\sum_{i=1}^{m^{++}} \frac{i}{i+m^{+-}} + \sum_{i=m^{+-}+1}^{m-k} \frac{m^{++}}{i+m^{++}}\right. \\ &\left. + \sum_{i=m^{++}+1}^{k} \frac{i}{i+m-k}\right]. \end{aligned}$$

As a direct consequence we see that if a ranking function $f$ performs well at a single precision-recall point (given by $m^{++}$ and $m^{+-}$), then the average precision *must* be within some region $\left[A_f^{\mathrm{MIN}}(C), A_f^{\mathrm{MAX}}(C)\right]$.

Simple examples of these curves are shown in Figure *3*. On the left, we show the case of an an accurate ranking function; for this system the possible choice of curves is very limited, and they all result in high average precision. On the contrary, on the right we have shown the same curves for a system at lower recall; in this case the range of attainable average precisions is much larger.

Note that although it may seem that we do not use any properties of the ranking function to establish these bounds, we do, in fact, use a crucial one namely its performance on a single point of the precision-recall curve. This method allows us to extend to the entire precision-recall curve whatever bounds




$q_{C,h} = r_{C,h} = 0.93$ $\qquad$ $q_{C,h} = 0.88,\ r_{C,h} = 0.47$

*Figure - 3*: Maximum and minimum possible average precision curves (see text for details).

we may have on a single point of this curve, and hence *to the average precision of the classifier*. For example, in Figure *4* we show a precision-recall curve on which we have established bounds on a single point, and we have shown how the maximum and minimum curves can be used to extend such a bound.

## 4. CONCLUSION

We have attempted to formalise the problem of generalisation in information retrieval. As IR systems gain in complexity, it becomes crucial to address this problem. Otherwise, little to no advancement can be made in the development of inference algorithms that directly optimise the performance measures of interest. We have considered McDiarmid-type bounds, and we proved the convergence of average precision. Thus we provided the building blocks of a learning theoretical treatment of information retrieval.

Furthermore, we have presented bounds on the maximum and minimum curves realisable given a particular precision-recall point of the curve. These bounds allow us to generalise existing bounds on single precision-recall points to the entire curve and hence to average-precision.

## ACKNOWLEDGEMENTS

We are greatly indebted to “Bob” Williamson for many valuable discussions.

## A. THE MAXIMUM AVERAGE PRECISION CHANGE

When considering how changing a single point can change the average precision of a precision-recall plot eight possible changes must be considered,

1. Ranking a positive document higher than previously.
2. Ranking a positive document lower than previously.
3. Ranking a positive document higher than previously and now denoting it as a negative one.

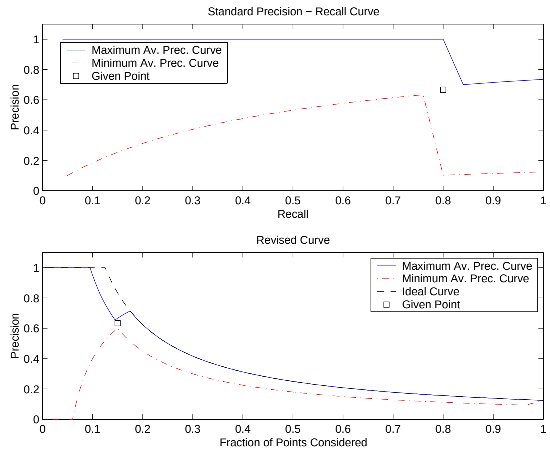



*Figure - 4*: Maximum and minimum possible average precision curves for a bounded precision-recall point.

4. Ranking a positive document lower than previously and now denoting it as a negative one.
5. Ranking a negative document higher than previously.
6. Ranking a negative document lower than previously.
7. Ranking a negative document higher than previously and now denoting it as a positive one.
8. Ranking a negative document lower than previously and now denoting it as a positive one.

## A.1 Derivation of the Eight Possible Changes

The approach to find each of the eight average precision changes is much the same. For brevity only the derivation relating to the first change is given here, all the others follow *mutatis mutandis*.

In moving a positive document, originally ranked $k$ to a new ranking $q \leq k$, the new average precision of a precision-recall curve is,

$$A_f^{NEW}(C) = \frac{1}{m^+}\left[\sum_{j=1}^{q-1}\frac{j}{i_{(j)+}} + \frac{q}{i'_{(q)+}} + \sum_{j=q}^{k-1}\frac{j+1}{i_{(j)+}+1} + \sum_{j=k+1}^{m^+}\frac{j}{i_{(j)+}}\right] \quad (11)$$

where $i_{(q-1)+} < i'_{(q)+} \leq i_{(q)+}$. This increases with decreasing $q$, to see this consider two cases, $q_1$ and $q_2$ where $q_2 = q_1 + 1$ and $i'_{(q_2)+} > i'_{(q_1)+}$. This being so then it can be seen that $i'_{(q_1)^+} \leq i_{(q_1)^+}$ and $i_{(q_1)^+} + 1 \leq i'_{(q_2)^+}$, and hence,

$$\begin{aligned}
\frac{q_1}{i'_{(q_1)^+}} + \frac{q_1+1}{i_{(q_1)^+}+1} &\geq \frac{q_1}{i_{(q_1)^+}} + \frac{q_1+1}{i'_{(q_2)^+}} \\
\sum_{j=1}^{q_1-1} \frac{j}{i_{(j)^+}} + \frac{q_1}{i'_{(q_1)^+}} + \sum_{j=q_1}^{k-1} \frac{j+1}{i_{(j)^+}+1} &\geq \sum_{j=1}^{q_2-1} \frac{j}{i_{(j)^+}} + \frac{q_2}{i'_{(q_2)^+}} + \sum_{j=q_2}^{k-1} \frac{j+1}{i_{(j)^+}+1}.
\end{aligned}$$

With this result then,

$$A_f^{NEW,MAX}(C) = \frac{1}{m^+}\left[1 + \sum_{j=1}^{k-1} \frac{j+1}{i_{(j)^+}+1} + \sum_{j=k+1}^{m^+} \frac{j}{i_{(j)^+}}\right] \tag{12}$$

$$A_f^{NEW,MIN}(C) = A_f(C). \tag{13}$$

Note that as $i_{(k+1)^+} \geq i_{(k)^+} + 1$ then,

$$\begin{aligned}
\frac{k+1}{i_{(k)^+}+1} &\geq \frac{k+1}{i_{(k+1)^+}} \\
\sum_{j=1}^{k} \frac{j+1}{i_{(j)^+}+1} + \sum_{j=k+2}^{m^+} \frac{j}{i_{(j)^+}} &\geq \sum_{j=1}^{k-1} \frac{j+1}{i_{(j)^+}+1} + \sum_{j=k+1}^{m^+} \frac{j}{i_{(j)^+}}
\end{aligned}$$

and so $A_f^{NEW,MAX}(C)$ increases with $k$, giving,

$$A_f^{NEW,MAX}(C) = \frac{1}{m^+}\left[1 + \sum_{j=1}^{m^+-1} \frac{j+1}{i_{(j)^+}+1}\right] \tag{14}$$

i.e. the maximum change occurs when the last positive document is moved to be the first positive document. This change in average precision is,

$$\begin{aligned}
\Delta A_f(C) &= A_f^{NEW,MAX}(C) - A_f(C) \\
&= \frac{1}{m^+}\left[1 - \frac{m^+}{i_{(m^+)^+}} + \sum_{j=1}^{m^+-1}\left(\frac{j+1}{i_{(j)^+}+1} - \frac{j}{i_{(j)^+}}\right)\right] \\
&= \frac{1}{m^+}\left[1 - \frac{m^+}{i_{(m^+)^+}} + \sum_{i=1}^{m^+-1} \frac{i_{(j)^+} - j}{i_{(j)^+}(i_{(j)^+}+1)}\right].
\end{aligned}$$

It now remains to determine what the maximum of this is, that is, which values of $i_{(j)^+}$ maximise the expression? Clearly $i_{(m^+)^+} = m$ does. For the documents in the summation what is required is that,

$$\frac{i_{(j)^+} - j}{i_{(j)^+}(i_{(j)^+}+1)} \geq \frac{i_{(j)^+} + 1 - j}{(i_{(j)^+}+1)(i_{(j)^+}+2)}$$

and

$$\frac{i_{(j)^+} - j}{i_{(j)^+}(i_{(j)^+}+1)} \geq \frac{i_{(j)^+} - 1 - j}{i_{(j)^+}(i_{(j)^+}-1)}.$$

Considering initially the first expression,

$$\begin{aligned}
(i_{(j)^+} - j)(i+2) &\geq i_{(j)^+}(i_{(j)^+} + 1 - j) \\
i_{(j)^+}^2 - 2j - i_{(j)^+}j + 2i_{(j)^+} &\geq i_{(j)^+}^2 + i_{(j)^+} - i_{(j)^+}j \\
i_{(j)^+} &\geq 2j
\end{aligned}$$

and now the second,

$$\begin{aligned}
(i_{(j)^+} - j)(i_{(j)^+} - 1) &\geq (i_{(j)^+} + 1)(i_{(j)^+} - 1 - j) \\
i_{(j)^+}^2 - i_{(j)^+}j + j - i_{(j)^+} &\geq i_{(j)^+}^2 - i_{(j)^+} - i_{(j)^+}j + i_{(j)^+} - 1 - j \\
i_{(j)^+} &\leq 2j + 1.
\end{aligned}$$

So $i_{(j)^+}$ equals either $2j$ or $2j+1$. To find which gives the greater outcome try both. First, with $i_{(j)^+} = 2j$,

$$\begin{aligned}
\frac{i_{(j)^+} - j}{i_{(j)^+}(i_{(j)^+} + 1)} &= \frac{2j - j}{2j(2j+1)} \\
&= \frac{1}{2(2j+1)},
\end{aligned}$$

second, with $i_{(j)^+} = 2j + 1$,

$$\begin{aligned}
\frac{i_{(j)^+} - j}{i_{(j)^+}(i_{(j)^+} + 1)} &= \frac{2j + 1 - j}{(2j+1)(2j+2)} \\
&= \frac{1}{2(2j+1)}
\end{aligned}$$

and so the result is the same in both cases. With this outcome then,

$$\Delta A_f^{MAX}(C) = \frac{1}{m^+}\left[1 - \frac{m^+}{m} + \frac{1}{2}\sum_{j=1}^{m^+-1} \frac{1}{2j+1}\right] \tag{15}$$

provided that $m \geq 2(m^+ - 1) + 1$, that is, $m \geq 2m^+ - 1$. This is not a particularly onerous condition in information retrieval, where generally positive documents are far fewer in number than negative ones.

Following this procedure results in eight possible average precision changes,

$$\Delta A_{f,1}^{MAX}(C) = \frac{1}{m^+}\left[1 - \frac{m^+}{m} + \frac{1}{2}\sum_{i=1}^{m^+-1} \frac{1}{2i+1}\right]$$

$$\Delta A_{f,2}^{MAX}(C) = \frac{1}{m^+}\left[1 - \frac{m^+}{m} + \frac{1}{2}\sum_{i=1}^{m^+-1} \frac{1}{2i+1}\right]$$

$$\Delta A_{f,3}^{MAX}(C) = \frac{1}{m^+}\left[1 + \frac{1}{m^+ - 1}\sum_{i=2}^{m^+} \frac{m^+ - i}{i}\right]$$

$$\Delta A_{f,4}^{MAX}(C) = \frac{1}{m^+} - \frac{1}{m}$$

$$\Delta A_{f,5}^{MAX}(C) = \frac{1}{m^+}\sum_{i=1}^{m^+}\frac{1}{i+1}$$

$$\Delta A_{f,6}^{MAX}(C) = \frac{1}{m^+}\sum_{i=1}^{m^+}\frac{1}{i+1}$$

$$\Delta A_{f,7}^{MAX}(C) = \frac{1}{m^+ + 1}\left[1 + \frac{1}{m^+}\sum_{i=1}^{m^+}\frac{m^+ - i}{i}\right]$$

$$\Delta A_{f,8}^{MAX}(C) = \frac{1}{m^+ + 1} - \frac{1}{m}.$$

The maximum of these is found in section A.2.

## A.2 Determination of the Maximum Possible Change

To find the maximum average precision change of those given in the previous section first note that,

$$\Delta A_{f,1}^{MAX}(C) = \Delta A_{f,2}^{MAX}(C). \tag{16}$$

Similarly,

$$\Delta A_{f,5}^{MAX}(C) = \Delta A_{f,6}^{MAX}(C). \tag{17}$$

Now, consider the next most straightforward comparisons. $\Delta A_{f,4}^{MAX}(C) = \frac{1}{m^+} - \frac{1}{m}$ is trivially greater than $\Delta A_{f,8}^{MAX}(C) = \frac{1}{m^+ + 1} - \frac{1}{m}$, as

$$\frac{1}{m^+} > \frac{1}{m^+ + 1}.$$

Now note that,

$$\Delta A_{f,3}^{MAX}(C) = \frac{1}{m^+ - 1}\sum_{j=2}^{m^+}\frac{1}{j} \tag{18}$$

$$\Delta A_{f,7}^{MAX}(C) = \frac{1}{m^+ + 1}\sum_{j=1}^{m^+}\frac{1}{j}. \tag{19}$$

From these then

$$\begin{aligned}
\Delta A_{f,7}^{MAX}(C) - \Delta A_{f,5}^{MAX}(C) &= \sum_{j=1}^{m+}\left(\frac{1}{m^+ + 1}\frac{1}{j} - \frac{1}{m^+}\frac{1}{j+1}\right) \\
&= \sum_{j=1}^{m^+}\frac{m^+ - j}{m^+(m^+ + 1)j(j+1)} \\
&> 0
\end{aligned}$$

and so $\Delta A_{f,7}^{MAX}(C) > \Delta A_{f,5}^{MAX}(C)$. Additionally,

$$\begin{aligned}
\Delta A_{f,7}^{MAX}(C) \quad - \quad & \Delta A_{f,3}^{MAX}(C) \\
= \quad & \frac{1}{m^+ + 1} + \left(\frac{1}{m^+ + 1} - \frac{1}{m^+ - 1}\right) \sum_{j=2}^{m^+} \frac{1}{j} \\
= \quad & \frac{1}{m^+ + 1}\left[1 - \frac{2}{m^+ - 1}\sum_{j=2}^{m^+}\frac{1}{j}\right] \\
\geq \quad & 0, \text{ for } m^+ \geq 2,
\end{aligned}$$

and so $\Delta A_{f,7}^{MAX}(C) \geq \Delta A_{f,3}^{MAX}(C)$ for $m^+ \geq 2$. Consider now $A_{f,1}^{MAX}(C)$,

$$\begin{aligned}
\Delta A_{f,7}^{MAX}(C) \quad - \quad & \Delta A_{f,1}^{MAX}(C) \\
= \quad & \frac{1}{m^+ + 1}\sum_{j=1}^{m^+}\frac{1}{j} - \frac{1}{m^+}\left[1 - \frac{m^+}{m} + \sum_{j=1}^{m^+ - 1}\frac{1}{4j+2}\right] \\
= \quad & \frac{1}{m^+(m^+ + 1)} + \frac{1}{m} - \frac{1}{m^+} + \sum_{j=1}^{m^+ - 1}\left(\frac{1}{m^+ + 1}\frac{1}{j} - \frac{1}{m^+}\frac{1}{4j+2}\right) \\
= \quad & \frac{1}{m^+(m^+ + 1)} + \frac{1}{m} - \frac{1}{m^+} + \sum_{j=1}^{m^+ - 1}\frac{4m^+ j + 2m^+ - m^+ j - j}{m^+(m^+ + 1)j(4j+2)} \\
= \quad & \frac{1}{m} - \frac{1}{m^+ + 1} + \frac{1}{2m^+(m^+ + 1)}\sum_{j=1}^{m^+ - 1}\frac{(3m^+ - 1)j + 2m^+}{j(2j+1)} \\
= \quad & \frac{1}{m} + \frac{1}{m^+ + 1} \times \left[\frac{1}{2m^+}\sum_{j=1}^{m^+ - 1}\frac{(3m^+ - 1)j + 2m^+}{j(2j+1)} - 1\right] \\
> \quad & 0, \text{ for } m^+ \geq 2.
\end{aligned}$$

Hence $\Delta A_{f,7}^{MAX}(C) > \Delta A_{f,1}^{MAX}(C)$ for $m^+ \geq 2$. Finally, then, it remains to compare $\Delta A_{f,7}^{MAX}(C)$ and $\Delta A_{f,4}^{MAX}(C)$. In doing this assume that $m \gg m^+$, with this assumption it will be shown that, although an upper bound is effectively considered on $\Delta A_{f,4}^{MAX}(C)$, $\Delta A_{f,7}^{MAX}(C)$ is still generally greater. To see this consider,

$$\begin{aligned}
(m^+ + 1)(\Delta A_{f,7}^{MAX}(C) - \Delta A_{f,4}^{MAX}(C)) \quad = \quad & \sum_{j=1}^{m^+}\frac{1}{j} - \frac{m^+ + 1}{m^+} \\
> \quad & 0, \text{ for } m^+ \geq 2,
\end{aligned}$$

so, from all possibilities, $\Delta A_{f,7}^{MAX}(C)$ has been shown to be greatest.

## A.3 Maximum Normalised Average Precision Changes of the Revised Curve

For the normalised average precision of the revised curve, the same eight changes apply. The working follows the general approach outlined in section A.1, *exceptis excipiendis.* Consideration must also now be made of the normalising area; $A^*$. With this then possible maximum area changes are,

$$\Delta A'^{MAX}_{f,1}(C) = \frac{1}{A^* m}\left[\frac{m-m^+}{m} + \sum_{i=1}^{m-m^+}\frac{1}{i+1} + \sum_{i=m-m^++1}^{m-1}\frac{m-m^+}{i(i+1)}\right]$$

$$\Delta A'^{MAX}_{f,2}(C) = \frac{1}{A^* m}\left[\sum_{i=2}^{m-m^++1}\frac{1}{i} + \sum_{i=m-m^++2}^{m}\frac{m-m^+}{i(i-1)} + 1 - \frac{m^+}{m}\right]$$

$$\Delta A'^{MAX}_{f,3}(C) = \frac{A^*\sum_{i=1}^{m}\frac{1}{i} - \frac{1}{mm^+}\left(\sum_{i=2}^{m^+}\frac{i}{i+m^-} + 2\right)}{A^* m\left(A^* - \frac{1}{mm^+}\right)}$$

$$\Delta A'^{MAX}_{f,4}(C) = \frac{1 + \frac{1}{m} + \frac{1}{m^+}\sum_{i=1}^{m^-}\frac{m^+-1}{i+m^+-1} - \frac{1}{m^+} - A^*}{A^* m^2\left(A^* - \frac{1}{mm^+}\right)}$$

$$\Delta A'^{MAX}_{f,5}(C) = \frac{1}{A^* m}\left[\sum_{i=1}^{m^+}\frac{1}{i+1} + \sum_{i=m^++1}^{m-1}\frac{m^+}{i(i+1)} + \frac{m^+}{m}\right]$$

$$\Delta A'^{MAX}_{f,6}(C) = \frac{1}{A^* m}\left[\sum_{i=2}^{m^++1}\frac{1}{i} + \sum_{i=m^++2}^{m}\frac{m^+}{i(i-1)} + \frac{m^+}{m}\right]$$

$$\Delta A'^{MAX}_{f,7}(C) = \frac{1 - \frac{m^+}{m} + \sum_{i=2}^{m}\frac{1}{i} - \sum_{i=m^-+1}^{m}\frac{i-m^-}{i-1}\left[\frac{1}{i} + \frac{1}{mA^*(m^++1)}\right]}{m\left(A^* + \frac{1}{m(m^++1)}\right)}$$

$$\Delta A'^{MAX}_{f,8}(C) = \frac{\left(\frac{1}{m^++1} - \frac{1}{m}\right)}{m\left(A^* + \frac{1}{m(m^++1)}\right)}$$

## REFERENCES


[1] I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals, Series and Products.* Academic Press, Inc., fifth edition, 1994.

[2] R. Herbrich. *Learning Kernel Classifiers: Theory and Algorithms.* MIT Press, 2002.

[3] T. Joachims. *The Maximum-Margin Approach to Learning Text Classifiers Methods, Theory, and Algorithms.* PhD thesis, Universität Dortmund, 2000.

[4] C. McDiarmid. On the method of bounded differences. In *Survey in Combinatorics*, pages 148–188. Cambridge University Press, 1989.

[5] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, 1995.