Mathematical Derivation of Expected AP Under Random Ranking

1. Problem Statement

The Average Precision Metric

Average Precision (AP) for a ranking with M relevant items out of L total items is defined as:

$$\mathrm{AP} = \frac{1}{M} \sum_{k: y_k = 1} \mathrm{Prec@}k$$

where Prec@k = (number of relevant items in top k) / k.

The Central Question

What is E[AP] when items are ranked uniformly at random?

The Incorrect Intuition

One might guess that E[AP] = p = M/L (the prevalence), thinking that on average, precision at any rank equals the overall proportion of relevant items. This common assumption was first formally shown to be incorrect by Bestgen (2015).

The Correct Answer

$$\mathbb{E}[\mathrm{AP}] = \frac{1}{L}\left[\frac{M-1}{L-1}(L - H_L) + H_L\right]$$

where $H_L = \sum_{k=1}^{L} \frac{1}{k}$ is the L-th harmonic number.

This is strictly greater than p for finite L, converging to p only as L → ∞.

2. The Exchangeability Derivation (Step-by-Step)

Step 1: Focus on One Relevant Item

Consider a uniformly random ranking. Pick any one of the M relevant items and let R denote its rank. By symmetry of random ranking:

$$P(R = r) = \frac{1}{L} \text{ for } r \in \{1, 2, \dots, L\}$$

Step 2: Condition on Position

Given that our chosen relevant item appears at rank r, what is the expected precision at that rank?

The precision at rank r is:

$$\text{Prec@}r = \frac{\text{number of relevant items in first r positions}}{r} = \frac{X+1}{r}$$

where X is the number of OTHER relevant items among the first r-1 positions (the "+1" accounts for our chosen item at position r).

Step 3: Count Other Relevant Items

Given R = r, the first r-1 positions are a random sample of size r-1 from the remaining L-1 items (which contain M-1 relevant items). Therefore:

$$X \mid R = r \sim \text{Hypergeometric}(L-1, M-1, r-1)$$

The expected value of a hypergeometric distribution is:

$$\mathbb{E}[X \mid R = r] = \frac{(r-1)(M-1)}{L-1}$$

Step 4: Compute Expected Precision at Rank r

$$\mathbb{E}[\text{Prec@}r \mid R = r] = \mathbb{E}\left[\frac{X+1}{r} \mid R = r\right] = \frac{1}{r}\left(\frac{(r-1)(M-1)}{L-1} + 1\right)$$

Simplifying:

$$\mathbb{E}[\text{Prec@}r \mid R = r] = \frac{(r-1)(M-1)}{(L-1)r} + \frac{1}{r}$$

Step 5: Average Over All Positions

By the exchangeability principle, each of the M relevant items contributes equally to the expected AP. The contribution of our chosen item is:

$$\mathbb{E}[\text{contribution}] = \mathbb{E}[\text{Prec@}R] = \sum_{r=1}^{L} P(R = r) \cdot \mathbb{E}[\text{Prec@}r \mid R = r]$$

$$= \frac{1}{L}\sum_{r=1}^{L}\left(\frac{(r-1)(M-1)}{(L-1)r} + \frac{1}{r}\right)$$

Since AP is the average contribution over all M relevant items, and each contributes the same amount:

$$\mathbb{E}[\text{AP}] = \mathbb{E}[\text{contribution}]$$

Step 6: Recognize the Harmonic Sum

We need to evaluate:

$$\mathbb{E}[\text{AP}] = \frac{1}{L}\sum_{r=1}^{L}\left(\frac{(r-1)(M-1)}{(L-1)r} + \frac{1}{r}\right)$$

Split this into two parts:

$$= \frac{1}{L} \cdot \frac{M-1}{L-1}\sum_{r=1}^{L}\frac{r-1}{r} + \frac{1}{L}\sum_{r=1}^{L}\frac{1}{r}$$

For the first sum, note that:

$$\sum_{r=1}^{L}\frac{r-1}{r} = \sum_{r=1}^{L}\left(1-\frac{1}{r}\right) = L - \sum_{r=1}^{L}\frac{1}{r} = L - H_L$$

For the second sum:

$$\sum_{r=1}^{L}\frac{1}{r} = H_L$$

Step 7: Final Form

Substituting back:

$$\mathbb{E}[\text{AP}] = \frac{1}{L}\left[\frac{M-1}{L-1}(L-H_L) + H_L\right]$$

This is our main result.

3. Alternative Form: Prevalence Plus Correction

Algebraic Manipulation

Starting from the main result, we can rewrite it to make the bias explicit:

$$\mathbb{E}[\text{AP}] = \frac{1}{L}\left[\frac{M-1}{L-1}(L-H_L) + H_L\right]$$

Expand:

$$= \frac{1}{L} \cdot \frac{M-1}{L-1} \cdot L - \frac{1}{L} \cdot \frac{M-1}{L-1} \cdot H_L + \frac{H_L}{L}$$

$$= \frac{M-1}{L-1} - \frac{(M-1)H_L}{L(L-1)} + \frac{H_L}{L}$$

Combine the H_L terms:

$$= \frac{M-1}{L-1} + \frac{H_L}{L} - \frac{(M-1)H_L}{L(L-1)}$$

$$= \frac{M-1}{L-1} + \frac{H_L(L-1) - (M-1)H_L}{L(L-1)}$$

$$= \frac{M-1}{L-1} + \frac{H_L \cdot N}{L(L-1)}$$

Now we want to express this in terms of prevalence p = M/L. Through algebraic manipulation, we can show that:

$$\frac{M-1}{L-1} = \frac{M}{L} - \frac{N}{L(L-1)} = p - \frac{N}{L(L-1)}$$

Therefore:

$$\mathbb{E}[\mathrm{AP}] = \frac{M}{L} - \frac{N}{L(L-1)} + \frac{H_L \cdot N}{L(L-1)}$$

$$= p + \frac{N(H_L - 1)}{L(L-1)}$$

This shows that E[AP] equals the prevalence p plus a positive correction term.

Asymptotic Behavior

Using the asymptotic expansion $H_L \approx \log L + \gamma$ where γ is the Euler-Mascheroni constant:

$$\mathbb{E}[\mathrm{AP}] - p = \frac{N(H_L - 1)}{L(L-1)} \approx \frac{(1-p)\log L}{L} + O(1/L)$$

This shows the bias decreases as O(log L / L), converging to zero logarithmically slowly.

4. The Hypergeometric Method (Bestgen 2015)

The Formula

Bestgen (2015) first identified the error in the prevalence approximation and provided this exact calculation using a double summation:

$$\mathbb{E}[\mathrm{AP}] = \frac{1}{M} \sum_{i=1}^{M} \sum_{n=i}^{N+i} P_{\text{hyper}}(i; L, M, n) \cdot \left(\frac{i}{n}\right)^2$$

where $P_{\text{hyper}}(i; L, M, n)$ is the hypergeometric PMF: the probability of getting exactly i successes when drawing n items from a population of L items containing M successes.

Understanding the Logic

For the i-th relevant item to appear:

1. **Where can it appear?** At ranks n in {i, i+1, …, N+i}
   - At least at rank i (if all previous i-1 are relevant)
   - At most at rank N+i (if all N non-relevant items come first, then i relevant ones)
2. **What's the probability?** For the i-th relevant item to appear at rank n:
   - Need exactly i relevant items in the first n positions: $P_{\text{hyper}}(i; L, M, n)$
   - Given i relevant in n positions, probability the i-th is at position n: i/n
   - Combined: $P_{\text{hyper}}(i; L, M, n) \cdot (i/n)$
3. **What's the contribution?** When the i-th relevant item appears at rank n with i relevant items total:
   - Precision at rank n is i/n
   - Contribution to AP is (i/n)
   - Total contribution: $P_{\text{hyper}}(i; L, M, n) \cdot (i/n) \cdot (i/n)$

Verifying Equivalence

Both the harmonic formula and Bestgen's hypergeometric method yield identical results when computed exactly. This equivalence has been verified across numerous test cases, including Bestgen's original examples from his 2015 paper.

The key insight is that both methods capture the same underlying probabilistic structure:

- The harmonic method uses exchangeability and conditional expectations
- The hypergeometric method explicitly sums over all possible configurations

Despite their different approaches, they arrive at the same exact formula, confirming the correctness of the derivation.

5. Key Insights

Why Harmonic Numbers?

Harmonic numbers appear naturally because:

- Precision at rank k involves a 1/k term
- Summing over all possible ranks gives Σ(1/k) = H_L
- The "partial" sum Σ(r-1)/r transforms to L - H_L

The Bias Structure

The correction term N(H_L - 1)/[L(L-1)] reveals:

- Bias is proportional to N (number of negatives)
- Bias grows logarithmically with L (since H_L ~ log L)
- Bias vanishes as O(log L / L) asymptotically

When the Bias Matters Most

The relative error (E[AP] - p)/p is approximately (1-p)/p × log(L)/L, which is largest when:

1. **Low prevalence** (small p): The factor (1-p)/p becomes large
   - This factor grows as p approaches zero
   - At balanced prevalence (p = 0.5), the factor equals 1
   - For high prevalence, the factor becomes small
2. **Small samples** (small L): The log(L)/L term is larger
   - The bias term decreases logarithmically with sample size
   - Even for large L, some bias persists due to the log L factor

Practical Impact

The bias is most significant for:

- **Low prevalence scenarios**: The relative error grows as (1-p)/p
- **Small datasets**: Common in specialized domains with limited labeled data
- **Statistical testing**: Using prevalence instead of the exact formula inflates Type I error rates

This matters for:

- Statistical significance testing (using p as null hypothesis inflates Type I error)
- Comparing algorithms to random baselines
- Small datasets common in specialized domains