

Optimal Check Ordering Problem

Matteo Biggio

February 1, 2017

1 Problem Statement

We are given a set S of n checks, with indices ranging from 1 to n , each one with an associated cost $c_i > 0$ and a KO probability $0 \leq P_i \leq 1$.

Checks are meant to be applied to objects taken from a given ensemble \mathcal{O} . Every check, when applied to a particular object $o \in \mathcal{O}$, can result in a KO, in which case we say o is *invalidated* by i , or in a OK, in which case another check is applied, until o is finally invalidated by some check or all the checks of S are OK for o (in which case we say that o *passes* the checks in S).

The KO probability P_i for check i is calculated as the fraction of objects of \mathcal{O} for which check i results in a KO:

$$P_i = \frac{|\{o \in \mathcal{O} : o \text{ is invalidated by } i\}|}{|\mathcal{O}|} \quad (1)$$

We wish to find, given the check costs $\{c_i, i = 1, \dots, n\}$ and the check KO probabilities $\{P_i, i = 1, \dots, n\}$, a permutation p of the indices $1, \dots, n$ such that the average summed check cost is minimized

$$\begin{aligned} C_{i_1, i_2, \dots, i_n} &= P_{i_1} c_{i_1} + (1 - P_{i_1}) P_{i_2} [c_{i_1} + c_{i_2}] + \dots \\ &\dots + (1 - P_{i_1}) \dots (1 - P_{i_{n-1}}) P_{i_n} [c_{i_1} + \dots + c_{i_n}] + \\ &\quad + (1 - P_{i_1}) \dots (1 - P_{i_n}) [c_{i_1} + \dots + c_{i_n}] \end{aligned} \quad (2)$$

or a little more formally

$$C_{i_1, i_2, \dots, i_n} = \sum_{k=1}^n \left\{ \left[\prod_{h=1}^{k-1} (1 - P_{i_h}) P_{i_k} \right] \left[\sum_{p=1}^k c_{i_p} \right] \right\} + \left(\prod_{h=1}^n (1 - P_{i_h}) \right) \left(\sum_{p=1}^n c_{i_p} \right) \quad (3)$$

Each terms of the summation from $k = 1$ to n of equation (3) represents the total cost of a sequence of checks, given that the first check that returns KO is the k -th one in the permutation, weighted by the corresponding probability $(1 - P_{i_1}) \dots (1 - P_{i_{k-1}}) P_{i_k}$. The last term is the total cost of a sequence of checks that all result in OK, weighted by the corresponding probability $(1 - P_{i_1}) \dots (1 - P_{i_n})$.

With some algebra, the cost function (3) can be simplified to the following (much more convenient) form

$$C_{i_1, i_2, \dots, i_n} = \sum_{k=1}^n \left\{ \prod_{h=1}^{k-1} (1 - P_{i_h}) \right\} c_{i_k} \quad (4)$$

2 A Greedy Solution

Despite the problem looking complicated, we will show in this section that a permutation of the checks that minimizes the total average cost (4) can be obtained very simply in $O(n \log n)$ worst-case time by ordering the checks in non-increasing order of the ratio P_i/c_i . This intuitively sounds like a good greedy strategy: checks that have high KO probability and little cost should come first in the optimal sequence so that they invalidate a big fraction of objects as soon as possible with little cost. It's however non-trivial that the exact tradeoff between the cost and the KO probability has to be expressed as a ratio, and this is what we are going to prove.

Theorem 1. *The sequence obtained by ordering the checks in non-increasing order of the ratio P_i/c_i is an optimal sequence for the optimal check ordering problem.*

Proof. We will prove the theorem by showing that ordering the elements by pairwise swaps between adjacent elements starting from an optimal solution never increases the cost function. Suppose there is an optimal sequence that is not ordered in non-increasing order of P_i/c_i , and suppose (for simplicity) that this sequence corresponds to the identity permutation of the checks. Let the optimal cost attained with this sequence be C_{opt} . Then it means that there are at least two consecutive checks i and $i + 1$, with $1 \leq i \leq n - 1$, such that $P_i/c_i < P_{i+1}/c_{i+1}$. Suppose we construct another permutation of the checks by swapping these two, and indicate its resulting cost as $C_{i \leftrightarrow i+1}$.

We want to calculate the difference $C_{i \leftrightarrow i+1} - C_{opt}$ using equation (4) for the cost function. It is easy to see that the only two terms that are affected by the swap are the ones with $k = i$ and $k = i + 1$:

$$\begin{aligned} C_{i \leftrightarrow i+1} - C_{opt} = & \left\{ \prod_{h=1}^{i-1} (1 - P_h) \right\} c_{i+1} + \left\{ \prod_{h=1}^{i-1} (1 - P_h) \right\} (1 - P_{i+1}) c_i \\ & - \left\{ \prod_{h=1}^{i-1} (1 - P_h) \right\} c_i - \left\{ \prod_{h=1}^{i-1} (1 - P_h) \right\} (1 - P_i) c_{i+1} \end{aligned} \quad (5)$$

Collecting common terms and simplifying

$$C_{i \leftrightarrow i+1} - C_{opt} = \left\{ \prod_{h=1}^{i-1} (1 - P_h) \right\} (c_{i+1} P_i - c_i P_{i+1}) \quad (6)$$

The first term is a product of probabilities and is therefore always greater than or equal to zero. On the other hand, the second term is less than zero, because of the way we choose indices i and $i + 1$. Therefore the cost function can never increase by swapping two elements that are in increasing order of the ratio P_i/c_i . Since every sequence of numbers can be ordered by a finite number of pairwise swaps between adjacent elements (insertion-sort algorithm), we have proven our theorem. \square