Algorithm 1 Pick-k

1: **input:** set size k, universe of items $\mathcal{Y}$, α, β, $f(.)$, $g(.)$
2: $Y \leftarrow \{\phi\},\ n \leftarrow k,\ \alpha' \leftarrow \alpha,\ \beta' \leftarrow \beta$
3: **while** $n > 0$ **do**
4: $\quad M \leftarrow \left\{\arg\max_s \frac{\alpha'/n+f(s)}{\beta'/n+g(s)}\right\} \qquad (s \in \mathcal{Y} \setminus Y)$
5: $\quad$ If $|M| > n$, then keep any n elements in M and throw away the rest
6: $\quad Y \leftarrow Y \cup M$
7: $\quad \alpha' \leftarrow \alpha' + \sum_{m\in M} f(m)$
8: $\quad \beta' \leftarrow \beta' + \sum_{m\in M} g(m)$
9: $\quad n \leftarrow n - |M|$
10: **end while**
11: **output:** picked elements $Y \subseteq \mathcal{Y}$

We want to optimize the following objective (the X subscripts have been dropped in the interest of clarity):

$$h(Y) = \frac{\alpha + \sum_i f(y_i)}{\beta + \sum_i g(y_i)}, \qquad (1)$$

We want to prove that Algorithm 1 picks the optimal solution $Y^* = \arg\max_Y h(Y)$.

Proof of Optimality. For ease of exposition, assume that there are no ties in step 4 and each iteration adds only one element to Y (the proof can be easily extended to cover that case of multiple additions per iteration).

The algorithm maximizes Equation 1 by solving a sequence of sub-problems. Suppose that the set of elements $Y^*_{(k-n)} = \{y^*_1, \dots, y^*_{k-n}\}$ is known to belong to Y^*. Now, for any set $W \subseteq \mathcal{Y} \setminus Y^*_{(k-n)}$ such that W has exactly n elements, we use Equation 1 to get:

$$h(Y^*_{(k-n)} \cup W) = \frac{\alpha + \sum_{y\in Y^*_{(k-n)}} f(y) + \sum_{w\in W} f(w)}{\beta + \sum_{y\in Y^*_{(k-n)}} g(y) + \sum_{w\in W} g(w)} \qquad (2)$$

Let $W^* = \arg\max_W h(Y^*_{(k-n)} \cup W)$. Now, we can define the following subproblem:

Problem 1': *Given the 4-tuple $(Y^*_{(k-n)}, \alpha, \beta, n)$, find any one element $w^* \in W^*$.*

Next, we relate the solution of the sub-problem to the optimal solution Y^*.

LEMMA 1. $w^* \in W^* \Rightarrow w^* \in Y^*$

PROOF. Clearly, $Y^* = Y^*_{(k-n)} \cup W^*$ in order to maximize Equation 1. Since $w^* \in W^*$, we must have $w^* \in Y^*$. $\square$

LEMMA 2. *Problem $1'$ with the 4-tuple $(Y^*_{(k-n)}, \alpha, \beta, n)$ is equivalent to the 4-tuple*
$$\left(\{\phi\}, \alpha + \sum_{y\in Y^*_{(k-n)}} f(y), \beta + \sum_{y\in Y^*_{(k-n)}} g(y), n\right).$$

PROOF. This follows from the form of Equation 2. $\square$

The optimality of Algorithm 1 can now be proved in two stages. First, assuming the correctness of our solution to each sub-problem, we show that the *sequence* of sub-problems generated by steps 6-9 of our algorithm is correct. Second, we show that each sub-problem is solved correctly (step 4).

THEOREM 1. *Assuming that step 4 of Algorithm 1 correctly solves the sub-problem $(\{\phi\}, \alpha', \beta', n)$, the algorithm returns the optimal result Y^*.*

PROOF. We show that each iteration adds one new element of Y^* to Y; as there are k iterations and $|Y^*| = k$, the resulting Y must equal Y^*.

The proof is by induction on the number of iterations $k - n$. In the first iteration ($k - n = 0$), the sub-problem being solved by step 4 is given by $(\{\phi\}, \alpha, \beta, k)$, whose solution is a member of Y^* by Lemma 1.

Suppose the first $k - n$ iterations are correct, and yield $Y^*_{(k-n)} \subset Y^*$. By repeated applications of steps 6-9, we must have $\alpha' = \alpha + \sum_{y\in Y^*_{(k-n)}} f(y)$, and $\beta' = \beta + \sum_{y\in Y^*_{(k-n)}} g(y)$. Thus, the sub-problem for the next iteration, viz. $(\{\phi\}, \alpha', \beta', n)$, is equivalent to $(Y^*_{(k-n)}, \alpha, \beta, n)$ by Lemma 2. Once again, solving this will yield another element of Y^*, by Lemma 1. Hence, at the end of $k - n + 1$ iterations, the set Y contains $k - n + 1$ elements of Y^*; $Y = Y^*_{(k-n+1)}$, as desired. This completes the proof. $\square$

Next, we prove the correctness of step 4. Consider one particular iteration, where n elements are yet to be added. For every $y \in \mathcal{Y}$, define num$(y) = \alpha'/n + f(y)$, den$(y) = \beta'/n + g(y)$, and imp(y) = num(y)/den(y). We call these the numerator, the denominator, and the importance of element y respectively. We reuse this notation for sets $S \subseteq \mathcal{Y}$ as well: num$(S) = \sum_{y\in S} f(y)$, den$(S) = \sum_{y\in S} g(y)$, and imp(S) = num(S)/den(S). Next, we prove an inequality between imp(S) of any set S and its members, which uses the following fact:

$$\frac{a}{b} > \frac{c}{d} \Rightarrow \frac{a}{b} > \frac{a+c}{b+d} > \frac{c}{d} \qquad (b, d > 0) \qquad (3)$$

THEOREM 2. *Let $S = (s_1, \dots, s_n)$ be an ordered set of size $n \geq 2$ such that imp(s_1) < imp(s_2) < … < imp(s_n). Then, imp(s_1) < imp(S) < imp(s_n).*

PROOF. We prove imp(s_1) < imp(S); the other case is similar. The proof is by induction on n.

When $n = 2$, imp$(S) = \frac{\text{num}(s_1)+\text{num}(s_2)}{\text{den}(s_1)+\text{den}(s_2)} > \frac{\text{num}(s_1)}{\text{den}(s_1)} =$ imp(s_1), using Fact 3.

Suppose the proposition is true for all sets of size $n - 1$. Then, for set S of size n, consider the subset $S' = S \setminus \{s_n\}$ of size $n - 1$. We have: imp$(S) = \frac{\text{num}(S)}{\text{den}(S)} = \frac{\text{num}(S')+\text{num}(s_n)}{\text{den}(S')+\text{den}(s_n)}$. Now, $\frac{\text{num}(S')}{\text{den}(S')}$ = imp(S') > imp(s_1), by assumption. Also, $\frac{\text{num}(s_n)}{\text{den}(s_n)}$ = imp(s_n) > imp(s_1), by the ordering of set S. From these, it is easily seen that imp(S) > imp(s_1). $\square$

Let $W^* = \{w^*_1, \dots, w^*_n\}$ be the optimal set of n elements that are yet to be selected in accordance with Equation 2, i.e., $Y^* = Y^*_{(k-n)} \cup W^*$. Without loss of generality, let imp(w^*_1) < imp(w^*_2) < … < imp(w^*_n)[1]. Now, note that step 4 of Algorithm 1 picks the element with the highest importance: $s^* = \arg\max_{s\in\mathcal{Y}\setminus Y}$ imp(s).

THEOREM 3. *Step 4 of Algorithm 1 correctly solves the sub-problem $(\{\phi\}, \alpha', \beta', n)$, i.e., the element selected in step 4 belongs to the optimal solution: $s^* \in W^*$.*

PROOF. We have:

$$\begin{aligned} h(Y^*) &= h(Y^*_{(k-n)} \cup W^*) \\ &= \frac{\alpha' + \sum_{i=1}^n f(w^*_i)}{\beta' + \sum_{i=1}^n g(w^*_i)} \\ &= \frac{\sum_{i=1}^n (\alpha'/n + f(w^*_i))}{\sum_{i=1}^n (\beta'/n + g(w^*_i))} \\ &= \text{imp}(W^*) \end{aligned} \qquad (4)$$

[1] Recall that ties are not being considered here, but can be easily folded into the proof.

This establishes the relationship between the function $h(.)$ that we want to optimize, and the function $\text{imp}(S)$ that we could express using the importances of individual elements $\text{imp}(s)$ where $s \in S$. This only works because the size of the set is n, and we have been using α'/n and β'/n in the imp(.) function.

Now, suppose $s^* \notin W^*$. Let $W' = W^* \setminus w_1^*$, and $Z = W' \cup \{s^*\} = \{w_2^*, \ldots, w_n^*, s^*\}$. Note that Z is again a set of n elements, and hence,

$$h(Y^*_{(k-n)} \cup Z) = \text{imp}(Z) = \frac{\text{num}(W') + \text{num}(s^*)}{\text{den}(W') + \text{den}(s^*)}.$$

Now,

$$\begin{aligned} \frac{\text{num}(W')}{\text{den}(W')} &= \text{imp}(W') \\ &< \text{imp}(w_n^*) \qquad \text{from Thm 2} \\ &< \text{imp}(s^*) \qquad \text{by definition of } s^* \end{aligned} \tag{5}$$

Similarly,

$$\begin{aligned} \frac{\text{num}(W')}{\text{den}(W')} &= \text{imp}(W') \\ &> \text{imp}(w_2^*) \qquad \text{from Thm 2} \\ &> \text{imp}(w_1^*) \end{aligned} \tag{6}$$

Thus, we have:

$$h(Y^*) = \text{imp}(W^*) = \frac{\text{num}(w_1^*) + \text{num}(W')}{\text{den}(w_1^*) + \text{den}(W')} < \frac{\text{num}(W')}{\text{den}(W')},$$

where we use Equation 6 and Fact 3. Similarly,

$$h(Y^*_{(k-n)} \cup Z) = \text{imp}(Z) = \frac{\text{num}(W') + \text{num}(s^*)}{\text{den}(W') + \text{den}(s^*)} > \frac{\text{num}(W')}{\text{den}(W')},$$

using Equation 5 and Fact 3. Hence,

$$h(Y^*) < \frac{\text{num}(W')}{\text{den}(W')} < h(Y^*_{(k-n)} \cup Z),$$

implying that Y^* is not the optimal solution, which is a contradiction. Hence, $s^* \in W^*$. $\square$