THE VECTOR SELECTION PROBLEM

LYNDON WHITE, WEI LIU

Definition 1. the Vector Selection problem is defined on ($\mathcal{V}$, $\tilde{s}$, d) for

- A finite vocabulary of vectors $\mathcal{V}$, $\mathcal{V} \subset \mathbb{R}^n$
- a target vector $\tilde{s}$, $\tilde{s} \in \mathbb{R}^n$
- any metric d

  by

$$\underset{\{\forall \tilde{c} \in \mathbb{N}_0^V\}}{\operatorname{argmin}} \; d(\tilde{s}, \sum_{j=1}^{j=V} \tilde{x}_j c_j)$$

- V is size of the vocabulary $\mathcal{V}$. (˜1300 for ATIS2, ˜50,000 for Brown, ˜10,000 for daily English)
- $\tilde{x}_j$ is the vector embedding for the jth word in the vocabulary $\tilde{x}_j \in \mathcal{V}$
  - We can express the embedding vocabulary $\mathcal{V} = \{\tilde{x}_j \,|\, \forall j \in \mathbb{N} \wedge 1 \leq j \leq V\} \subset \mathbb{R}^n$
  - If we treat $\mathcal{V}$ as a matrix with vectors $\tilde{x}_j$ for rows,(and treat length 1 vectors as scalars) we get the compact notation

$$\underset{\{\forall \tilde{c} \in \mathbb{N}_0^J\}}{\operatorname{argmin}} \; d(\tilde{s}, \sum_{j=1}^{j=V} \tilde{x}_j c_j) = \underset{\{\forall \tilde{c} \in \mathbb{N}_0^J\}}{\operatorname{argmin}} \; d(\tilde{s}, \mathcal{V}\, \tilde{c}^T)$$

- c_j is the count of how many times the jth word in the vocabulary occurs. $\tilde{c} \in \mathbb{N}_0^V$, so $c_j \in \mathbb{N}_0$
- n is the dimensional of the word vectors, $n = 300$ in current trials.

0.1. An Analogy for the problem.

(which may or may not help)

Imaging you are in a tile shop which as a variety of rectangular (2D) tiles. They have many copies of each tile (an unlimited number in-fact), but only a finite number of different sizes. This tiles have connectors on them, like jigsaw pieces, such that you can attach a North/South side to another North/South side even on a tile of different size, and similar for the East/West sides. But you can't attach a North/South side to a East/West side. i.e. You can not rotated the tiles.

You have 2 lengths given as your target when choosing tiles: a North/South length, and a East/West length.

Your task is to select a collection of tiles from the store, such that when connected on the north/south and east/west the total length in those directions is as close as possible to those to targets.

A formula is given for how your solution will be judged. It takes the form of some distance metric. E.g it might be the your distance from the target east/west length with your connected tiles, plus your distance from the target north/south length. Or maybe accuracy on north/south is twice as important as east/west. Or north/south difference squared etc. Your task it to minimize that score.

For example, if the store had 3 types of tiles. 4.1×4.1,1.5×5.0 and 100×1, and your targets were 13.1 and 20.0, and the scoring was Manhattan.

you might choose one 4.1×4.1 tiles and three 1.5×5.0 tiles, giving you length totals 8.6 and 19.1 and a score of 5.4

Had you chosen to take an extra 4.1×4.1 tile though given totals of 12.7 and 23.2 giving a better score of 4.5

Now generalize it from 2D tiles to hyperblocks of some arbitrary dimensionality.

0.2. Reduction from Subset sum.

The subset sum problem is well known to be NP-complete. First shown in by Karp under the name "Knapsack"[1] which has since come to be used for the more general problem.

It can be defined with the question: for a given set $\mathcal{S} \subset \mathbb{Z}$, does there exists $\mathcal{L} \subseteq \mathcal{S}$ such that $\sum_{l_i \in \mathcal{L}} l_i = 0$?

We reduce from subset sum to the Vector Selection Problem by showing any general solution to the Vector Selection Problem could be used to solve subset sum with only linear time additional work.

Claim 2. Any method which can solve the Vector Selection Problem will allow Subset sum to be completed with only linear time additional operations

- Let $\mathcal{S} = \{w_1, w_2, ..., w_m\}$
- Let $\Omega = 2m \left(\max_{i \in [1,m]} |w_i| + 1\right)$ and thus larger than the largest possible sum of elements of $\mathcal{S}$.

    - Finding this is a linear time operation, the only such operation in this method.
- Let $\omega = \frac{1}{2m}$ and thus smaller than any element of $\mathcal{S}$, except if $0 \in \mathcal{S}$ (in which case the solution is trivial)
- then we can define an embedding vocabulary $\mathcal{V}_s$ from based on $\mathcal{S}$ by

$$\mathcal{V}_s = \{[[w_i,\, 1]\,;\, \hat{e}_i] \,:\, w_i \in \mathcal{S}\}$$

    - By imposing some arbitrary total ordering on $\mathcal{S}$.
    - where ; is the concatenation operator,
    - and $\hat{e}_i$ is the elementary basis unit vector for dimension i. ie a vector with all zeros, except at index i, where it is 1.
    - i.e. we take the image of $\mathcal{S}$ into $\mathcal{V}_S$, by the function:

$$w_i \mapsto \begin{bmatrix} w_i \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \hat{e}_{i+2}$$

    - In doing so we map each integer w_i in $\mathcal{S}$ to a point in $\mathbb{R}^{m+1}$ where
        * the first index is the integer, w_i,
        * the second a term is used to force a solution that is nonempty to be better than an empty solution – all other things being equal;
        * the remaining m terms are used to force a solution which uses the same element more than once to be worse than one which uses it once or zero times.
    - Note that $\mathcal{V}_S \subset \mathbb{R}^{m+2}$
- we define the target vector by $\tilde{s_s} = \left[[0,m]\,;\, 0.5\sum_{j=1}^{j=m}\ \hat{e}_j\right] = \begin{pmatrix} 0 \\ m \\ 0.5 \\ \vdots \\ 0.5 \end{pmatrix}$
- we define the distance metric being given by a weighed Manhattan distance (i.e. weighted L1 Norm).

$$d_s\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}\right) = |x_1 - y_1| + \omega\,|x_2 - y_2| + \Omega \sum_{j=3}^{j=n} |x_j - y_j|$$

    - We will prove that d_s is a metric below.
- The procedure for using these once defined to solve subset sum is:
    - the Vector Selection Problem for $(\mathcal{V}_s, \tilde{s}_s, d_s)$ is solved getting back $\tilde{c}^\star$
    - if $\tilde{c}^\star = \mathbf{0}$, or $\sum_{j=1}^{j=m}\ \tilde{x}_{j,1} c_j^\star \neq 0$ then no such solution exists, otherwise:
    - such a subset $\mathcal{L} \subset \mathcal{S}$ does exist, and is given by $\mathcal{L} = \{w_i \in \mathcal{S} \,:\, c_i^\star \geq 1\}$.
    - Note that it does not matter if $c_i^\star > 1$ as for such cases clipping it to multiplicity 1 is just as optimal. Which we will prove below.

Proof. d_s is a metric

The d_s is a special case of

$$d\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}\right) = \sum_{1 \leq j \leq n} \omega_i\, d'\left(x_j, y_j\right)$$

for d' a metric defined on scalars and $\forall i$ where $\omega_i, x_i, w_i \in \mathbb{R}$ and $\omega_i > 0$

Below it is shown that that this is always a metric, and thus d_s is a metric, by showing it meets the 3 requirements: of the coincidence axiom, being symmetric, and of the triangle inequality. The following properties hold for all $\tilde{a}, \tilde{b}, \tilde{c} \in \mathbb{R}^n$.

If d follows the coincidence axiom: $d(x,y) = 0 \iff x = y$. This is shown by:
if $\tilde{a} = \tilde{b}$ then $\forall j \in [1,n]\ d'(a_j, b_j) = 0$ thus $d(\tilde{a}, \tilde{b}) = 0$.

and if $d(\tilde{a},\tilde{b})=0$ then as all $\forall j \in [1,n]$ $w_j > 0$, therefore $d'(a_j,b_j)=0$.

If d is symmetric then $d(x,y)=d(y,x)$. Shown by:

$$d(\tilde{a},\tilde{b}) = \sum_{j=1}^{j=n} \omega_i \, d'\left(a_j,b_j\right) = \sum_{j=1}^{j=n} \omega_i \, d'\left(b_j,a_j\right) = d(\tilde{b},\tilde{a})$$

Thus d is symmetric

if d follows the triangle inequality then $d(x,z) \leq d(x,y)+d(y,z)$.
making use of $d'\left(a_j,b_j\right) + \, d'\left(b_j,c_j\right) \geq d'(a_j,c_j)$, It is shown:

$$\begin{aligned}
d(\tilde{a},\tilde{b})+d(\tilde{b},\tilde{c}) &= \sum_{1\leq j\leq n} \omega_i \, d'\left(a_j,b_j\right) + \sum_{1\leq j\leq n} \omega_i \, d'\left(b_j,c_j\right) \\
d(\tilde{a},\tilde{b})+d(\tilde{b},\tilde{c}) &= \sum_{1\leq j\leq n} \omega_i \left(d'\left(a_j,b_j\right) + \, d'\left(b_j,c_j\right)\right) \\
d(\tilde{a},\tilde{b})+d(\tilde{b},\tilde{c}) &\leq \sum_{1\leq j\leq n} \omega_i \left(d'(a_j,c_j)\right) \\
d(\tilde{a},\tilde{b})+d(\tilde{b},\tilde{c}) &\leq d(\tilde{a},\tilde{b})
\end{aligned}$$

Thus d is a metric, and so d_s is a metric. □

Proof. Show for $c_j \notin \{0,1\}$ a better or at least equally good solution can be found for $c_j^{alt} \in \{0,1\}$

For a proof by contraction, we assume the existence of some optimal solution the Vector Selection Problem, c^* where for at least one index i,$c_i^* \geq 2$.

Consider also some alternative count vector (which by our assumption, can not be more optimal)

$$c' = c^* - \hat{e}_i$$

That is, c' is the same as c^*except that there is one less count for c_i^*

recalling $\tilde{s}_s = \begin{pmatrix} 0 \\ m \\ 0.5 \\ \vdots \\ 0.5 \end{pmatrix}$

we define the optimal sum of vectors by $\tilde{t}^*$

$$\tilde{t}^* = \sum_{j=1}^{j=V} \tilde{x}_j c_j^* = \begin{bmatrix} \left(\sum_{j=1}^{j=m} w_j c_j^*\right) \\ \left(\sum_{j=1}^{j=m} w_j c_j^*\right) \\ c_1^* \\ \vdots \\ c_n^* \end{bmatrix}$$

we define the alternative sum of vectors by $\tilde{t}'$

$$\tilde{t}' = \sum_{j=1}^{j=V} \tilde{x}_j c'_j = \left(\sum_{j=1}^{j=V} \tilde{x}_j c^*_j\right) - \tilde{x}_i = \begin{bmatrix} \left(\sum_{j=1}^{j=m} w_j c^*_j\right) - w_i \\ \left(\sum_{j=1}^{j=m} w_j c^*_j\right) - 1 \\ c^*_1 \\ \vdots \\ c^*_{i-1} \\ c^*_i - 1 \\ c^*_{i+1} \\ \vdots \\ c^*_n \end{bmatrix}$$

Note: we know that $c^*_i > 0.5$ as $c^*_i \geq 2$. Similarly we know $c'_i \geq 0.5$ for the as $c'_i = c^*_i - 1$

$$d_s(\tilde{s}_s, \tilde{t}^*) = \begin{array}{cc} & \left(\sum_{j=1}^{j=m} w_j c^*_j\right) \\ + & \omega \left| \left(\sum_{j=1}^{j=m} c^*_j\right) - m \right| \\ + & \Omega \left| c^*_1 - 0.5 \right| \\ & \vdots \\ + & \Omega \left| c^*_{i-1} - 0.5 \right| \\ + & \Omega (c^*_i - 0.5) \\ + & \Omega \left| c^*_{i+1} - 0.5 \right| \\ & \vdots \\ + & \Omega \left| c^*_n - 0.5 \right| \end{array}$$

and

$$d_s(\tilde{s}_s, \tilde{t}') = \begin{array}{cc} & \left(\sum_{j=1}^{j=m} w_j c^*_j\right) - w_i \\ + & \omega \left| \left(\sum_{j=1}^{j=m} c^*_j\right) - 1 - m \right| \\ + & \Omega \left| c^*_1 - 0.5 \right| \\ & \vdots \\ + & \Omega \left| c^*_{i-1} - 0.5 \right| \\ + & \Omega (c^*_i - 0.5 - 1) \\ + & \Omega \left| c^*_{i+1} - 0.5 \right| \\ & \vdots \\ + & \Omega \left| c^*_n - 0.5 \right| \end{array}$$

Since $\tilde{t}^*$ from the more optimal solution: $d_s(\tilde{s}_s, \tilde{t}') - d_s(\tilde{s}_s, \tilde{t}^*) \geq 0$

$$\begin{array}{cc} & \left(\sum_{j=1}^{j=m} w_j c^*_j\right) - w_i \\ + & \omega \left| \left(\sum_{j=1}^{j=m} c^*_j\right) - 1 - m \right| \\ + & \Omega \left| c^*_1 - 0.5 \right| \\ & \vdots \\ + & \Omega \left| c^*_{i-1} - 0.5 \right| \\ + & \Omega (c^*_{i+1} - 0.5 - 1) \\ + & \Omega \left| c^*_{i+1} - 0.5 \right| \\ & \vdots \\ + & \Omega \left| c^*_n - 0.5 \right| \end{array} - \begin{array}{cc} & \left(\sum_{j=1}^{j=m} w_j c^*_j\right) \\ + & \omega \left| \left(\sum_{j=1}^{j=m} c^*_j\right) - m \right| \\ + & \Omega \left| c^*_1 - 0.5 \right| \\ & \vdots \\ + & \Omega \left| c^*_{i-1} - 0.5 \right| \\ + & \Omega (c^*_{i+1} - 0.5) \\ + & \Omega \left| c^*_{i+1} - 0.5 \right| \\ + & \vdots \\ + & \Omega \left| c^*_n - 0.5 \right| \end{array} \geq 0$$

And after canceling terms:

$$-w_i + \omega \left(\left| \left(\sum_{j=1}^{j=m} c^*_j\right) - 1 - m \right| - \left| \left(\sum_{j=1}^{j=m} c^*_j\right) - m \right| \right) - \Omega \geq 0$$

let $K = \left(\sum_{j=1}^{j=m} c^*_j\right) - m$

$$-w_i + \omega\left(|K-1| - |K|\right) - \Omega \geq 0$$

The largest value $|K-1| - |K|$ can take is 1. (The other cases are 0, and -1, both of which result in the contradiction of the sum of 2 and 3 negative values respectively being greater than or equal to zero)

$$-w_i + \omega - \Omega \geq -w_i + \omega\left(|K-1| - |K|\right) - \Omega \geq 0$$

$$w_i + \Omega \leq \omega$$

Substituting in the values from the definitions:
$\Omega = 2m\left(\max_{j\in[1,m]} |w_j| + 1\right)$ and $\omega = \frac{1}{2m}$

$$w_i + 2m\left(\max_{j\in[1,m]} |w_j|\right) + 2m \leq \frac{1}{2m}$$

$$2mw_i + 4m^2\left(\max_{j\in[1,m]} |w_j|\right) + 4m^2 \leq 1$$

Assume w_i takes the most negative value possible: $w_i = -\left(\max_{j\in[1,m]} |w_j|\right)$ giving:

$$\left(4m^2 - 2m\right)\left(\max_{j\in[1,m]} |w_j|\right) + 4m^2 \leq 2mw_i + 4m^2\left(\max_{j\in[1,m]} |w_j|\right) + 4m^2 \leq 1$$

As $m \geq 1$, consider it taking that the smallest value it can take (so $m = 1$)
$2\left(\max_{j\in[1,m]} |w_j|\right) + 4 \leq \left(4m^2 - 2m\right)\left(\max_{j\in[1,m]} |w_j|\right) + 4m^2 \leq 1$
requiring, $\max_{j\in[1,m]} |w_j| \leq -\dfrac{3}{2}$
Which is impossible as the absolute value of an integer is always non-negative.
Thus a contradiction.
Thus c' is at least as optimal as c^*.
We may apply this proof to all claimed optimal solutions with a $c_i > 1$ to show that an equally optimal (or more so), solution has that c_i at 1 lower.
Thus if some solution with any count $c_i > 1$ is found, it can be transformed into a solution that is equally (or more so) optimal, by clipping all counts c_i at one. □

A finer proof could be developed showing strict inequality and that c' yields a strictly better solution that c^*

Proof. Proof of Correctness
let $\tilde{c}'$ be the solution to the Vector Selection Problem on $(\mathcal{V}_s, \tilde{s}_s, d_s)$
As it was shown above that for any $c'_i > 1$ an equally optimal solution can be created by clipping $c'_i$to 1.
We will thus assume $c'_i \in \{0, 1\}$.
let $L' = \{w_i \in \mathcal{S} \, : \, c'_i = 1\}$

Case 1. Subset Sum Exists, but the Vector Selection Problem based method says it does not
We assume for a proof by contradiction that the the Vector Selection Problem based method states that no such subset sub exists,
however it is incorrect and such a subset does and is given by $L^* \subseteq \mathcal{S}$.

Then $\mathcal{L}^*$defines a indicator vector $c^* \in \{0,1\}^m$, given by $c^*_j = \begin{cases} 1 & w_j \in \mathcal{L}^* \\ 0 & w_j \notin \mathcal{L}^* \end{cases}$, where w_j is the jth element of $\mathcal{S}$.
so $\sum_{j=1}^{j=m} w_j c^*_j = 0$.
Note also as $\mathcal{L}^* \neq \emptyset$ (by definition of subset sum) $\exists i \in [1,m]$ such that $c^*_i = 1$.
we define $\tilde{t}^*$ to be the sum of the vectors which correspond to c^*_j by

$$\tilde{t}^* = \sum_{j=1}^{j=V} \tilde{x}_j c_j^* = \begin{bmatrix} \sum_{j=1}^{j=m} w_j c_j^* \\ \sum_{j=1}^{j=m} c_j^* \\ c_1^* \\ \vdots \\ c_m^* \end{bmatrix} = \begin{bmatrix} 0 \\ \sum_{j=1}^{j=m} c_j^* \\ c_1^* \\ \vdots \\ c_m^* \end{bmatrix}$$

and so

$$d_s(\tilde{s},\,\tilde{t}^*) = d_s\left(\begin{bmatrix}0\\m\\0.5\\\vdots\\0.5\\\vdots\\0.5\end{bmatrix}, \begin{bmatrix}0\\\sum_{j=1}^{j=m} c_j^*\\c_1^*\\\vdots\\c_m^*\end{bmatrix}\right) = \begin{array}{cc} & 0\\ + & \omega\left|\left(\sum_{j=1}^{j=m} c_j^*\right)-m\right|\\ + & \Omega\,|c_1^*-0.5|\\ & \vdots\\ + & \Omega\,|c_n^*-0.5|\end{array} = \omega\left|\left(\sum_{j=1}^{j=m} c_j^*\right)-m\right| + 0.5m\Omega$$

since $0 < \sum_{j=1}^{j=m} c_j^* \leq m$ as it is sum of m variables $0 \leq c_j^* \leq 1$ and not all $c_j^* = 0$, we can that to simplify to

$$d_s(\tilde{s},\,\tilde{t}^*) = \omega\left(m - \sum_{j=1}^{j=m} c_j^*\right) + 0.5m\Omega$$

Now then consider the cases when the method (incorrectly) reports no such subset exists:

Case i. $\tilde{c}' = [0, ..., 0]$ (the zero vector)
We define the total sum of vectors given by $\tilde{t}'$

$$\tilde{t}' = \sum_{j=1}^{j=V} \tilde{x}_j c_j' = \begin{bmatrix} \sum_{j=1}^{j=m} w_j c_j' \\ \sum_{j=1}^{j=m} c_j' \\ c_1' \\ \vdots \\ c_m' \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

So find

$$d_s(\tilde{s},\,\tilde{t}') = Multidimentionald_s\left(\begin{bmatrix}0\\m\\0.5\\\vdots\\0.5\\\vdots\\0.5\end{bmatrix}, \begin{bmatrix}0\\0\\0\\\vdots\\0\end{bmatrix}\right) = \omega m + 0.5m\Omega$$

thus $d_s(\tilde{s},\,\tilde{t}^*) < d_s(\tilde{s},\,\tilde{t}')$ and so $c' = [0, ..., 0]$ could not have been the solution returned for solving the Vector Selection Problem as it is not the correct selection for the argmax. Thus a contradiction.

Case ii. $\tilde{c}' \neq \mathbf{0}$ thus $\sum_{j=1}^{j=m} \tilde{x}_{j,1} c_j' = k$ for $k \neq 0$
So thus the method reports that there is no nonempty subset which sums to zero; (the closest it can get is summing to k.
We redefine $\tilde{t}'$ for this case to be

$$\tilde{t}' = \sum_{j=1}^{j=V} \tilde{x}_j c_j' = \begin{bmatrix} \sum_{j=1}^{j=m} w_j c_j' \\ \sum_{j=1}^{j=m} c_j' \\ c_1' \\ \vdots \\ c_m' \end{bmatrix} = \begin{bmatrix} k \\ \sum_{j=1}^{j=m} c_j' \\ c_1' \\ \vdots \\ c_m' \end{bmatrix}$$

and so

$$d_s(\tilde{s},\,\tilde{t}') = d_s(\begin{bmatrix} 0 \\ m \\ 0.5 \\ \vdots \\ 0.5 \\ \vdots \\ 0.5 \end{bmatrix}, \begin{bmatrix} k \\ \sum_{j=1}^{j=m} c'_j \\ c'_1 \\ \vdots \\ c'_m \end{bmatrix}) = \begin{array}{cc} & |k| \\ + & \omega \left| \left(\sum_{j=1}^{j=m} c'_j \right) - m \right| \\ + & \Omega \left| c'_1 - 0.5 \right| \\ & \vdots \\ + & \Omega \left| c'_n - 0.5 \right| \end{array} = |k| + \omega \left| \left(\sum_{j=1}^{j=m} c'_j \right) - m \right| + 0.5m\Omega$$

As $c'$was selected over c^* then
by our
recalling:

$$d_s(\tilde{s},\,\tilde{t}^*) = \omega \left| \left(\sum_{j=1}^{j=m} c^*_j \right) - m \right| + 0.5m\Omega$$

as $\tilde{t}'$ is the sum of vectors giving min value for the distance to the target point $\tilde{s_s}$ thus

$$d_s(\tilde{s_s},\,\tilde{t}') \leq d_s(\tilde{s_s},\,\tilde{t}^*)$$

i.e.

$$|k| + \omega \left| \left(\sum_{j=1}^{j=m} c'_j \right) - m \right| + 0.5m\Omega \leq \omega \left| \left(\sum_{j=1}^{j=m} c^*_j \right) - m \right| + 0.5m\Omega$$

let $C' = \left(\sum_{j=1}^{j=m} c'_j \right)$ and $C^* = \left(\sum_{j=1}^{j=m} c^*_j \right)$
as $0 \leq C' \leq m$ and $0 \leq C^* \leq m$ as both are sums of indicator variables (0,1)
$\left| \left(\sum_{j=1}^{j=m} c^*_j \right) - m \right| = |C^* - m| = m - C^*$ and similarly for C'
substituting in:
$|k| + \omega\,(m - C') + 0.5m\Omega \leq \omega\,(m - C^*) + 0.5m\Omega$
i.e $|k| + \omega C' \leq \omega C^*$
i.e $|k| \leq \omega\,(C^* - C')$
thus $(C^* > C')$ it can not be equal as otherwise $k = 0$ which would be a contradiction.
let $(C^* - C') = C_d,\, C_d \in \mathbb{N}$
$0 < C_d$ as otherwise $|k| = 0$ (which would be the contradiction)
$C_d \leq m$ as the largest case is $C^* = m$ and $C' = 0$
Substitute

$$|k| \leq \omega C_d$$

substitute from the definition of $\omega = \frac{1}{2m}$

$$|k| \leq \frac{1}{2m} C_d$$

consider the largest value C_d can take: $C_d = m$

$$|k| \leq \frac{m}{2m}$$

$$|k| \leq \frac{1}{2}$$

As k is an integer this would mean $k = 0$
But this is a contradiction as $|k| > 0$.
Therefore it is not possible for the the Vector Selection Problem based method to say there is no solution if there is a solution.
i.e. if a solution exists, the the Vector Selection Problem based method will find it.

Case 2. Case: Subset sum does not exists, but the Vector Selection Problem based method says it does

Then:

$$\sum_{j=1}^{j=m} \tilde{x}_{j,1} c'_j = 0$$

.

We know that $\tilde{c}' \neq \mathbf{0}$ as other wise the the Vector Selection Problem based method would have said no solution exists.

Thus $\mathcal{L}' \neq \emptyset$

further we know by definition of $\tilde{x_j}$ that $\tilde{x}_{j,1} = w_j$ for $w_j \in \mathcal{S}$

thus we have $\sum_{j=1}^{j=m} w_j c'_j = 0$

thus in fact the sum of the elements of $\mathcal{L}'$ is zero.

And so a subset sum does exist.

This is a contradiction, thus the the Vector Selection Problem based method will never say there is a solution unless one exists.

Thus the method described in Claim 2 is a correct method to solve subset sum.

□

0.2.1. *Subset Sum Reduction Concluding note:* Thus it has been shown that if a general solution to the vector selection problem can be found a solution to subset sum could be found which would take at most a linear amount of additional time. Thus were a polynomial time solution for the Vector Selection Problem found, it would show that $P = NP$. However, the proof above is only for the general case, which is defined over ($\mathcal{V}$, $\tilde{s}$, d) for finite subsets of $\mathbb{R}^n$, $\mathcal{V}$; and any$\tilde{s} \in \mathbb{R}^n$, using any metric d. Thus the hardness result is only for the general case. Like for many problems from the knapsack family, there certainly exists special cases for which faster solutions are possible. For example $\mathcal{V} \subset \mathbb{R}^1_+$, $\tilde{s} = [0]$ and $d = (x, y) \mapsto |x - y|$, a linear time solution exists, found by finding the index of the smallest member of $\mathcal{V}$. The general problem however is not expected to have an exact solution in polynomial time.

References

1. Richard M Karp, *Reducibility among combinatorial problems*, Springer, 1972.