Finding a Maximum Weight Sequence with Dependency Constraints

M.Sc. Essay

Behrooz Sepehry

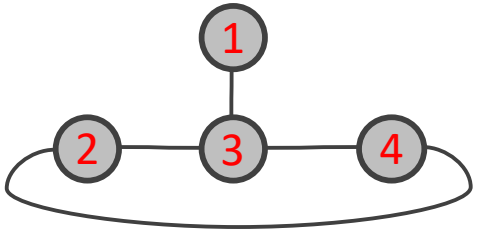
behrooz.sepehry@gmail.com

The Main Problem

- We are given a graph $G = (V, E)$, a weight $W(v_i)$ associated with each vertex $v_i \in V$, and an iteration number T. We want to choose a sequence of vertices $\boldsymbol{V} = \{v_{i_t}\}_{t=1}^{T}$ that maximizes $\sum W(v_{i_t})$, subject to some constraints.
- The constraints: After each time a vertex v_i is selected, it cannot be selected again until after a neighbor of vertex v_i has been selected.

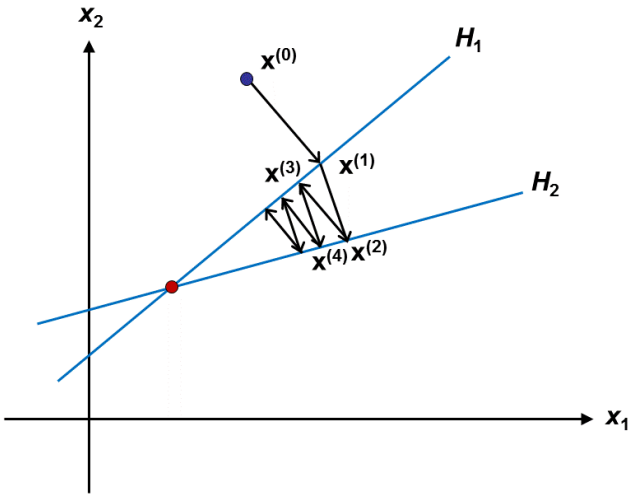
The Main Problem Cont.

- We call the problem Maximum weight Sequence with Dependency constraints or MSD.
- Example: $\boldsymbol{V}_1 = (2,2,2)$ is invalid
- Example: $\boldsymbol{V}_2 = (2,4,1,2,4,1)$ is invalid
- Example: $\boldsymbol{V}_3 = (3,1,4,3,1,4)$ is valid



Motivation: Greedy Kaczmarz

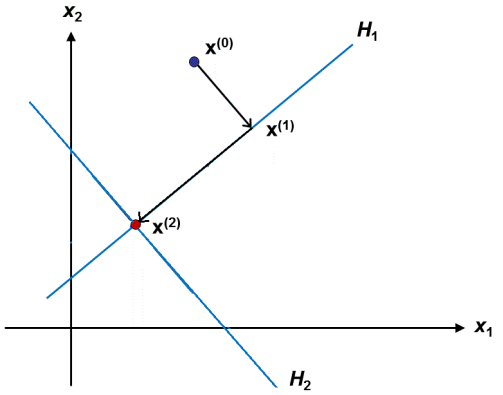
- Large scale linear systems of equations is a fundamental problem in machine learning.
- Kaczmarz method is an iterative method to solve the linear systems of equations $Ax = b$.
- At each step t, it selects a row i_t according to "some rule" and projects x^t onto the hyper plane corresponding to the row, i.e. the hyper plane $a_i x^t = b_i$.
- Random, cyclic and greedy are some possible selection rules.
- Kaczmarz method is faster than conjugate gradient when the number of rows is large.



Kaczmarz method
Source:
https://ccjou.wordpress.com/2014/01/06/kaczmarz-%E7%AE%97%E6%B3%95/

Motivation: Greedy Kaczmarz Cont.

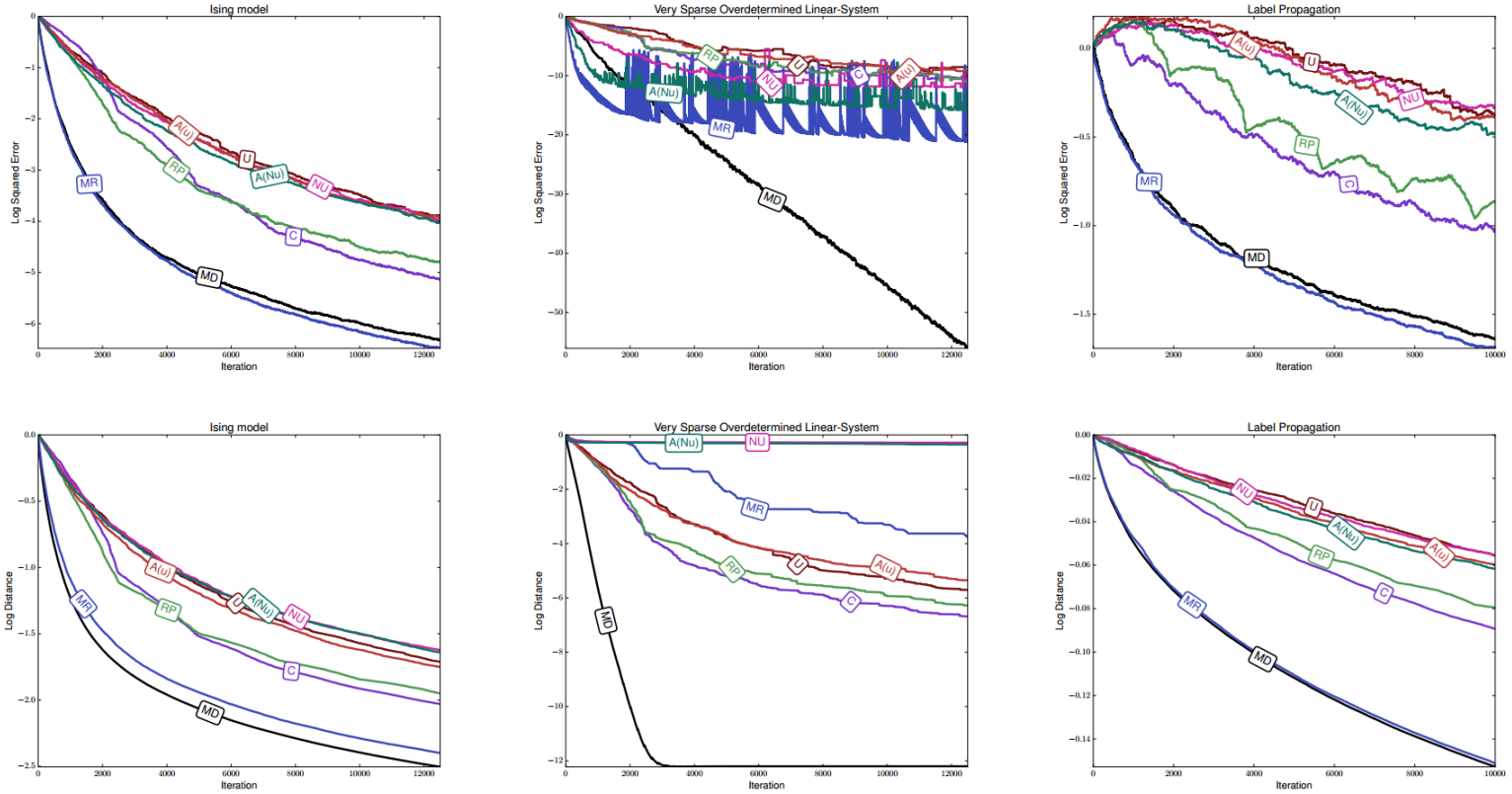
- Maximum residual selection rule: $i_t = \underset{i}{\operatorname{argmax}} |a_i^\top x^t - b_i|.$
- Can be implemented with the same cost as randomized Kaczmarz if the matrix A is sparse.
- If two hyper planes are orthogonal, projecting onto one will not change the residual of the other.
- We construct a graph $G = (V, E)$, such that for each row i we have a vertex v_i and there is an edge between two vertices i and j iff the corresponding rows are not orthogonal
- Because of the selection rule, after a row is selected, it will not be selected again until after a neighbor of it has been selected.



Motivation: Greedy Kaczmarz, Convergence Rate

- Convergence rate:
$$||x^t - x^*||^2 \leq \exp\left[\sum_{t=1}^{T} \ln\left(1 - \frac{\sigma(A,\infty)}{||a_{i_t}||^2}\right)\right] ||x^0 - x^*||^2,$$
- A trivial worst case analysis is $\exp\left(T \ln\left(1 - \frac{\sigma(A,\infty)}{\max ||a_i||^2}\right)\right)$. But in practice, the greedy rule is usually much faster than other selection rules, hinting that this convergence rate may not be tight.
- To find a better upper bound on the convergence rate, we can solve the MSD problem with graph $G = (V, E)$ and weight function $W(v_i) = \ln\left(1 - \frac{\sigma(A,\infty)}{||a_i||^2}\right)$.
- We can use a very similar approach to find a tighter bound on the convergence rate of coordinate descent with Gauss-Southwell rule with exact optimization.

Motivation: Greedy Kaczmarz, Convergence Rate Cont.

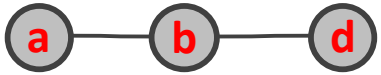


Comparison of Kaczmarz selection rules for squared error (top) and distance to solution (bottom).

Source: Convergence Rates for Greedy Kaczmarz Algorithms, and Faster Randomized Kaczmarz Rules Using the Orthogonality Graph, Nutini et al.

Solution

- Cyclical sequence: A sequence that can be repeated indefinitely.



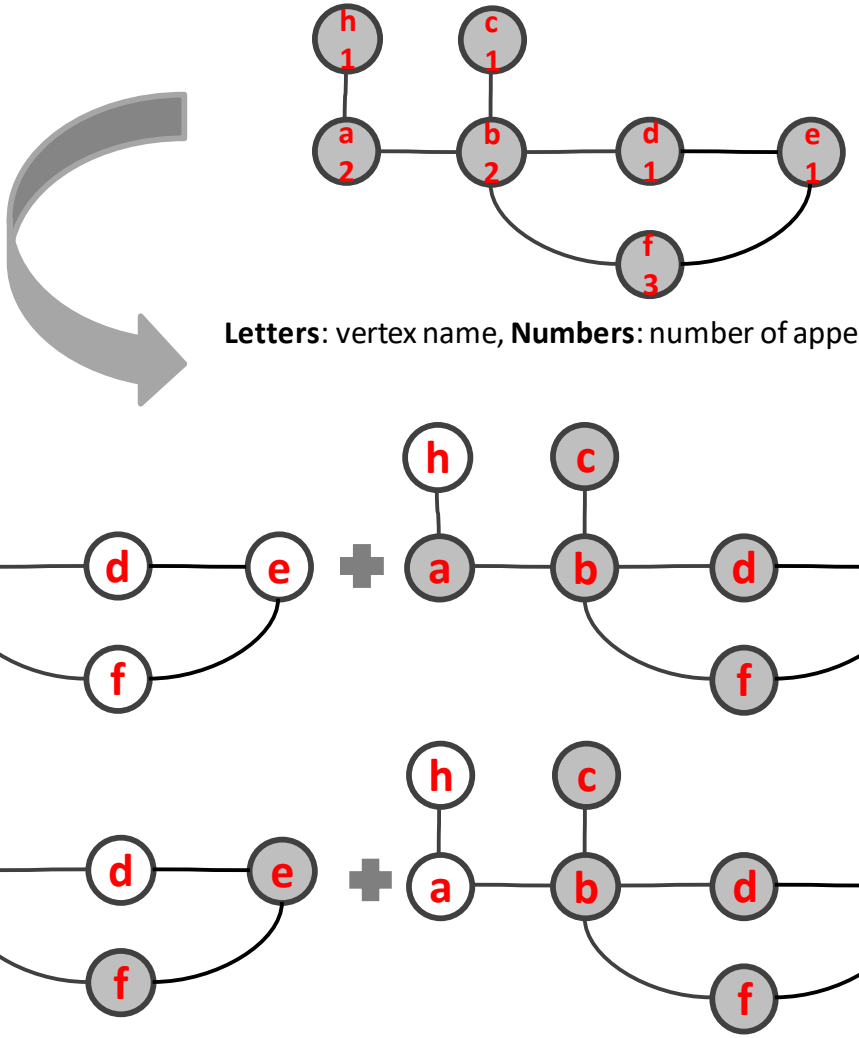
- Example: (a, b) is cyclical
- Example: (a, d) is not cyclical
- To approximate a long sequence with the highest average, we can repeat a cyclical sequence with the highest average. We can prove that when the length goes to infinity, the average weight of the two sequences would be the same.
- So to find an asymptotic solution, we can find a cyclical sequence with the highest average.

Solution Cont.

- We can prove that we can always decompose any valid cyclical sequence into several smaller cyclical sequences in which every selected vertex appears once and the sub-graph corresponding to the sequence is a star sub-graph.
- The decomposition is not necessarily unique, but always exists.

Solution Cont. Example

- Consider the sequence $\boldsymbol{V} = (a, h, a, b, c, d, f, b, f, e, f)$. Note that it is cyclical.
- We can decompose it to star sub-graphs.
- Star graphs has cyclical sequences.
- We can decompose $\boldsymbol{V}$ to $(a, h) + (b, a, c, d, f) + (b, c, d, f) + (e, f)$.
- Notice that edge decomposition doesn't work!



Letters: vertex name, **Numbers**: number of appearance

Grey: Selected, **White**: Not selected

Solution Cont.

- Among the cyclical sequences with the highest average, at least one of them should be a sequence in which every selected vertex appears once and the corresponding sub-graph is a star sub-graph.
- Because if not, then we can decompose the cyclical sequence with the highest average, and find one.
- To find a cyclical sequence with highest average, we only need to search over the star sub-graphs. Using this fact, we can find the sequence in $O(|V| \log|V| + |E|)$.

Solution Cont.

- The asymptotic upper bound on the convergence rate resulted from the solution of the MSD problem is:

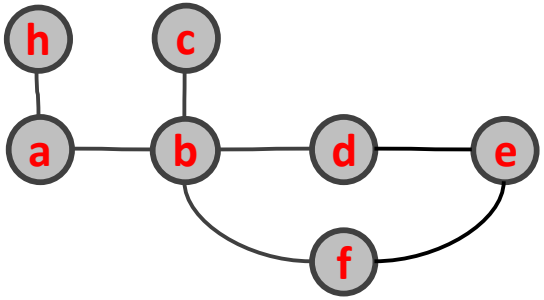
$$\|x^k - x^*\|^2 \leq$$

$$O(1)\left(\max_{S(G)}\left\{\sqrt[|S(G)|]{\prod_{j\in S(G)}\left(1-\frac{\sigma(A,\infty)^2}{\|a_j\|^2}\right)}\right\}\right)^T R_0^2,$$

- It can be much faster than the trivial bound $\left(1-\frac{\sigma(A,\infty)}{\max\left||a_i|\right|^2}\right)^T$ if rows with large $\left||a_i|\right|$ are not adjacent in the graph G.

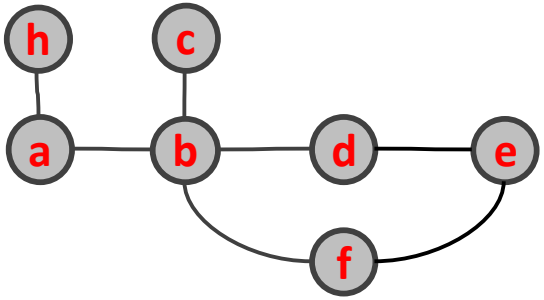
Generalizations: k-times-MSD

- k-times-MSD: similar to MSD, but after a neighbor of a vertex is selected, the vertex can be selected for k times instead of once.
- We can show that k-times-MSD can be reduced to MSD in $O\left(k^2(|V| + |E|)\right)$ time.



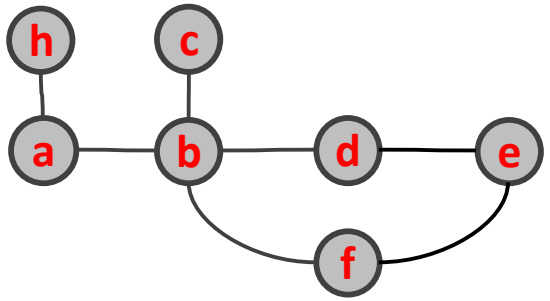
Generalizations: k-order-MSD

- k-order-MSD: similar to MSD, but after a vertex is selected, the vertices in distance k of the selected vertex become selectable.
- We can show that k-order-MSD can be reduced to MSD in $O(|V|^2)$ time.



Generalizations: k-neighbors-MSD

- k-neighbors-MSD: similar to MSD, but after a vertex is selected, we need to select at least k neighbors of the vertex to be able to select it again.
- For $k = 1$, the k-neighbors-MSD is the original MSD.
- For $k \geq 3$ We can prove that k-order-MSD is NP-Hard.
- For $k = 2$, I do not know!



Generalizations: probabilistic-SD

- probabilistic-SD: similar to MSD with the same constraints, but the vertices are selected randomly.
- probabilistic-SD can be modeled with a Markov chain.
- We can find the expected average weight of the sequence when the length goes to infinity from the steady state of the Markov chain.
- But the size (number of states) of the Markov chain is exponential.

Conclusion

- When trying to find the convergence rates for greedy Kaczmarz method and Coordinate descent with Gauss-Southwell rule, we can use solve the MSD problem to find a tight upper bound on the convergence rate.
- We found an asymptotic solution for the MSD problem, which results in an asymptotic bound on the convergence rates, which is tighter than previous ones, explaining why the methods work well in practice.
- We considered several generalizations of the MSD problem. Some of them were easy, and some of them were hard.

Motivation: Coordinate descent with Gauss-Southsell rule

- Coordinate descent with Gauss-Southsell rule and exact optimization wants to find $\min h(x)$:

$$h(x) = \sum_{i \in V} g_i(x_i) + \sum_{(i,j) \in E} f_{i,j}(x_i, x_j),$$

- Selection rule:

$$i_t = \underset{i}{\operatorname{argmin}} \, |\nabla_i h(x^t)|$$

$$\alpha_t = \underset{\alpha}{\operatorname{argmin}} \{h(x^\iota + \alpha e_{i_t})\}$$

$$x^{t+1} = x^t + \alpha_t e_{i_t}.$$

- Because of this selection rule, after a coordinate is selected, it cannot be selected again until after a neighbor of it has been selected.
- Convergence rate:

$$h(x^t) - h(x^*) \leq \exp\left[\sum_{t=1}^{T} \ln\left(1 - \frac{\mu_1}{L_{i_t}}\right)\right] [h(x^0) - h(x^*)],$$

- To find an upper bound on the convergence rate, we can solve the MSD problem with graph $G = (V, E)$ and weight function $W(i) = \ln\left(1 - \frac{\mu_1}{L_i}\right)$.