CONVERGENCE OF THE STOCHASTIC IHT ALGORITHM

Anonymous authors
Paper under double-blind review

ABSTRACT

The Iterative Hard Thresholding (IHT) algorithm has been considered extensively as an effective deterministic algorithm for solving sparse optimizations. The IHT algorithm benefits from the exact information of the gradient at each point and this information is a crucial key for the convergence analysis of the generated sequence. However, this strength becomes a weakness when it comes to the machine learning and high dimensional statistical applications. In these applications the objective function is defined as a summation of many individual functions each of which depends solely on one data point. Hence, calculating of the gradient is computationally infeasible. In this paper, we approximate the gradient with a mini-batch gradient and study the Stochastic IHT algorithm for solving the sparse optimizations. By finding a lower bound on the mini-bach gradient size, we establish a critical stochastic gradient descent property. Using this stochastic gradient descent property, we show that the sequence of the function values is a supermartingale sequence and converges. We test the SIHT algorithm on the mean square error and logistic losses to elaborate on our theoretical results.

1 INTRODUCTION

An optimization problem with cardinality constraint is known to be an NP-hard problem. The goal of this paper is to show this is not the case for the quadratic loss. To show this we have run ... million experiments. For better understanding consider $\mathbf{A} \in \mathbb{R}^{4\times 4}$ and $\mathbf{b} \in \mathbb{R}^4$ that are random. Then we get a quadratic function as $f(\mathbf{x}) = \frac{1}{2(4)}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$ where 4 is because we have 4 rows that determines that number of samples. Each function has $\binom{4}{2}$. This problem has at most 6 local minimizers. We want to answer why the below frequency table is formed when we start from random initialization.

conv_label	
1	116
2	518
3	554
4	687
6	125

Figure 1: Frequency of the number of convergence to different local minimizers

- Why certain basic stationary minimizers are accessible than others?
  1. Does it depend on the function value for the local minimizer?
  2. Does it depend on the absolute value of the non-supported gradient?
  3. Does it depend on the absolute value of the least nonzero element of the local minimizer?
  4. Does it depend on the gap that is defined by the formula of the unstable-stationary point?
  5. What is the average number of convergence for all local minimizers?
  6. What is the average number of steps for convergence to a specific local minimizer?

7. What are the eigenvalues of sub-matrix associated with a local minimizer?

- Why some of them are not accessible at all? Because are unstable-stationary point.
- Is there any way to always catch the best one which is global minimizer?
- Using features that we have found so far to classify local minimizers?
- Is there any local minimizer with positive gap whose frequency is zero? What would be the characteristic of this point? (seed number is 117952). However, there are sample with very small gap whose probability is very small (seed number is 439107).
- If we increase the variance of the initial condition, would the frequency of convergence change?

2 ANALYSIS

The process that we have done so far is the following. We pick a random seed, once we have that random seed, we get $\mathbf{A} \in \mathbb{R}^{4\times 4}$ and $\mathbf{b} \in \mathbb{R}^4$ that are random. Then we get a quadratic function as $f(\mathbf{x}) = \frac{1}{2(4)}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$ where 4 is because we have 4 rows that determines that number of samples. Each function has $\binom{4}{2}$.

x_1	x_2	x_3	x_4	x_L2	x_L1	f_value
-0.39	0.26	0.00	0.00	0.47	0.65	0.53
-0.41	0.00	-1.35	0.00	1.41	1.75	0.10
-0.66	0.00	0.00	1.26	1.42	1.92	0.25
0.00	0.60	-1.50	0.00	1.62	2.10	0.05
0.00	0.14	0.00	0.78	0.80	0.93	0.54
0.00	0.00	-1.27	0.37	1.33	1.65	0.19

Figure 2: Description of the local minimizers

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

In descriptions add which indices and what local mimizers.

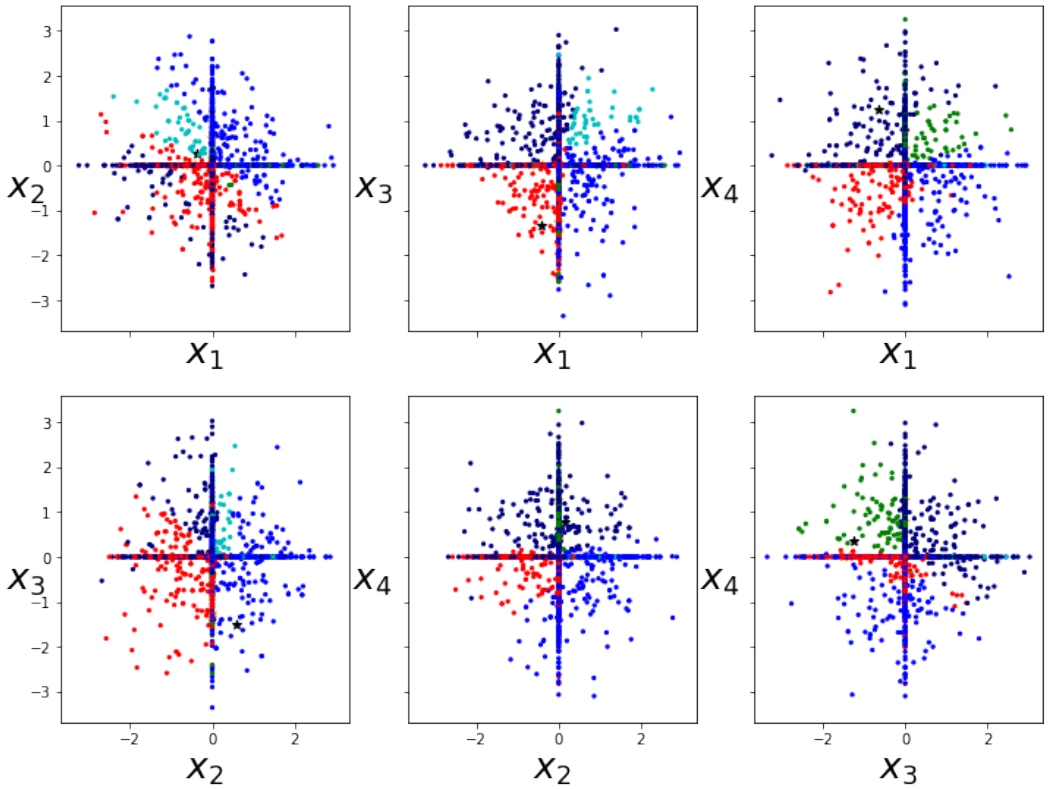

Figure 3: All the initial points for 2000 experiments

REFERENCES

A APPENDIX

You may include other additional sections here.

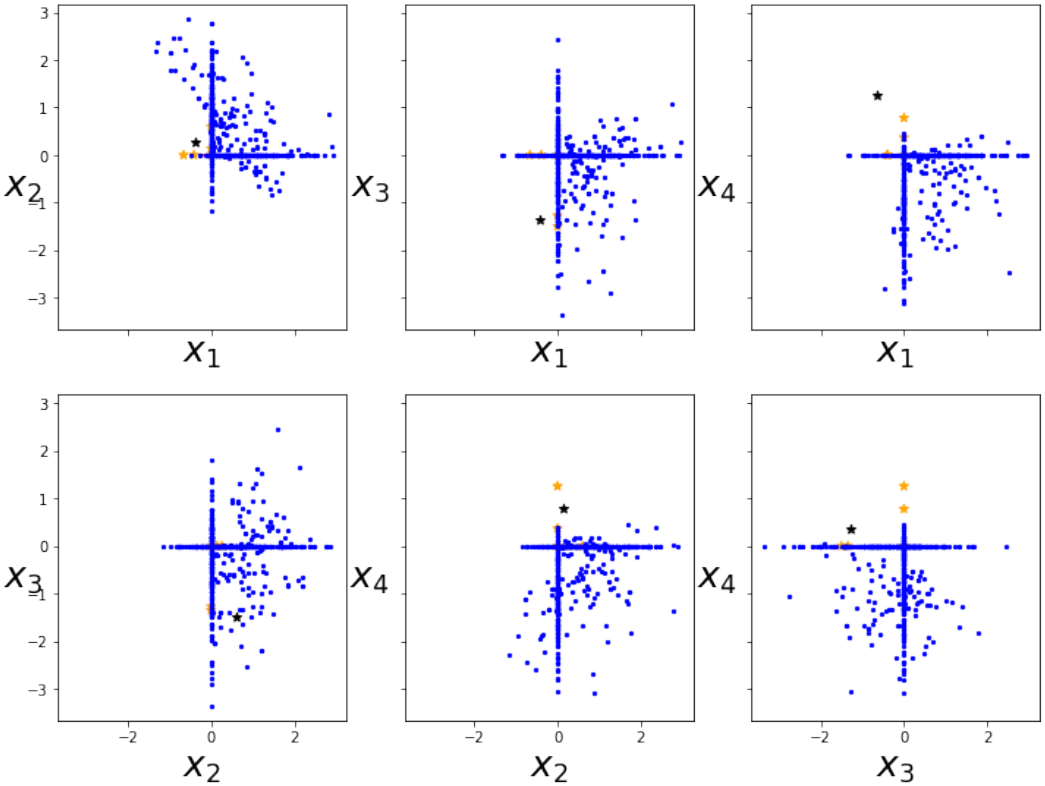

Figure 4: Initial points converging to $\mathbf{x}^{*(1)} = [0, 0.6, -1.5, 0]^\top$ whose support is $\text{supp}(\mathbf{x}^{*(1)}) = \{2, 3\}$ and the function value is $f^{*(1)} = 0.05$ (frequency out of 2000 experiments is 687 and seed=152,315).

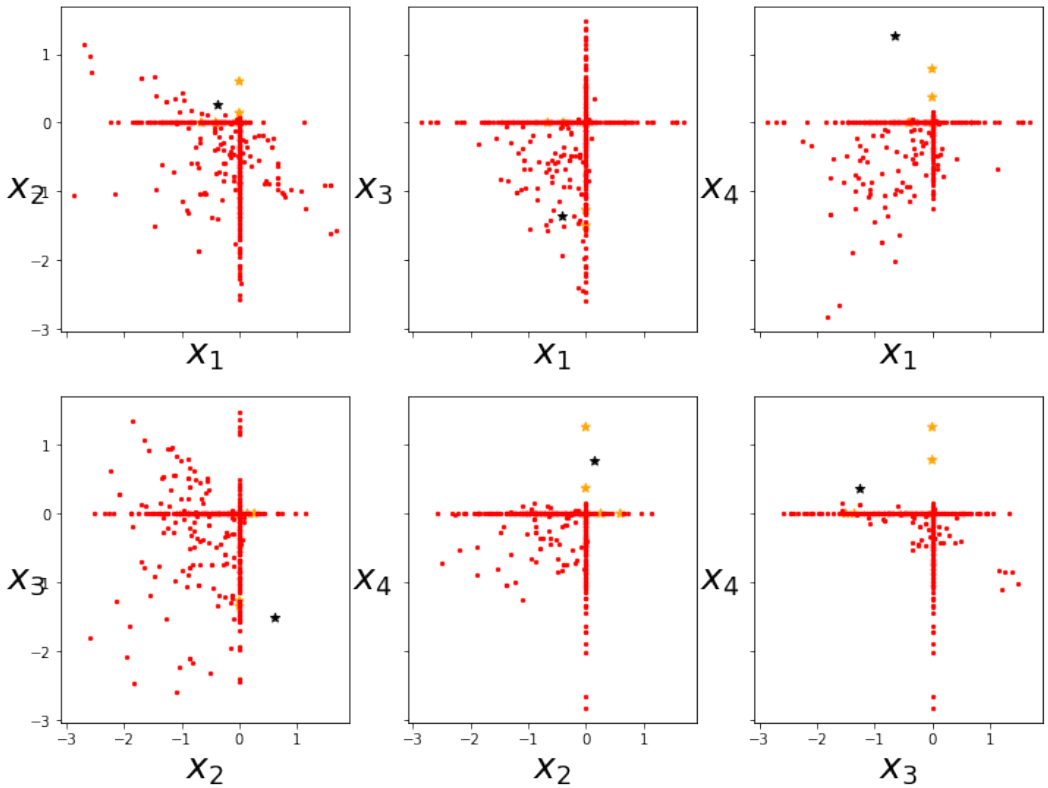

Figure 5: Initial points converging to $\mathbf{x}^{*(2)} = [-0.41, 0, -1.35, 0]^\top$ whose support is $\text{supp}(\mathbf{x}^{*(2)}) = \{1, 3\}$ and the function value is $f^{*(2)} = 0.1$ (frequency out of 2000 experiments is 518 and seed=152,315).

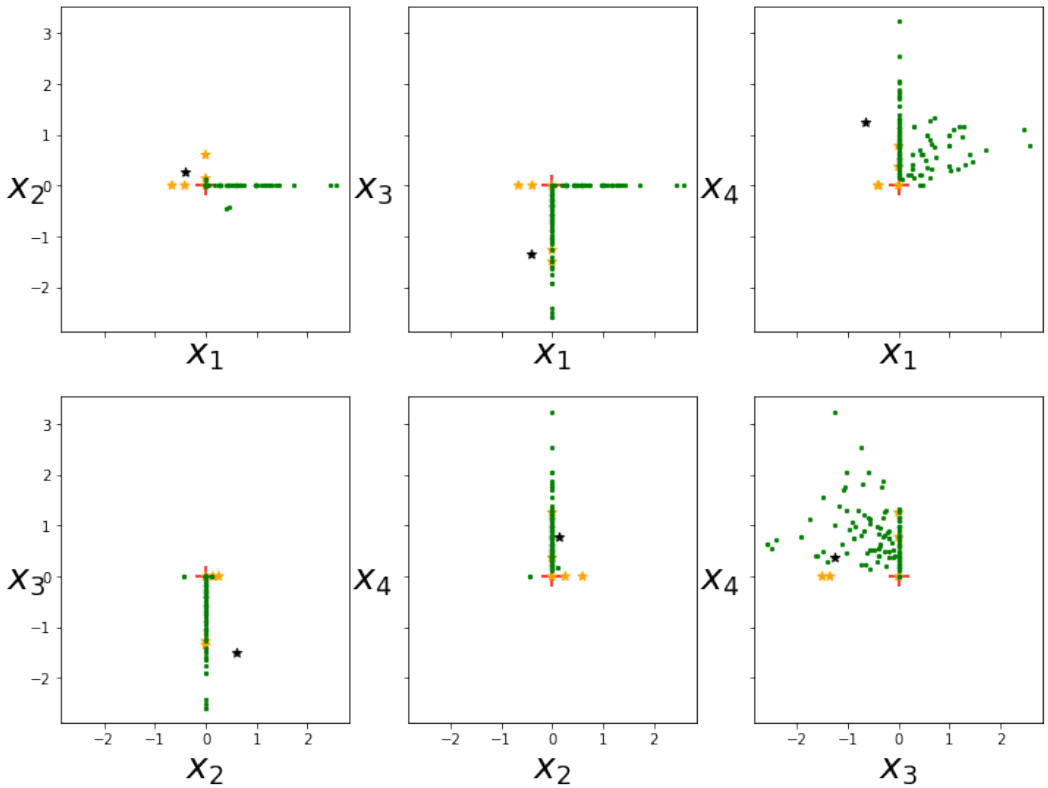

Figure 6: Initial points converging to $\mathbf{x}^{*(3)} = [0, 0, -1.27, 0.37]^\top$ whose support is $\text{supp}(\mathbf{x}^{*(3)}) = \{3, 4\}$ and the function value is $f^{*(3)} = 0.19$ (frequency out of 2000 experiments is 125 and seed=152,315).

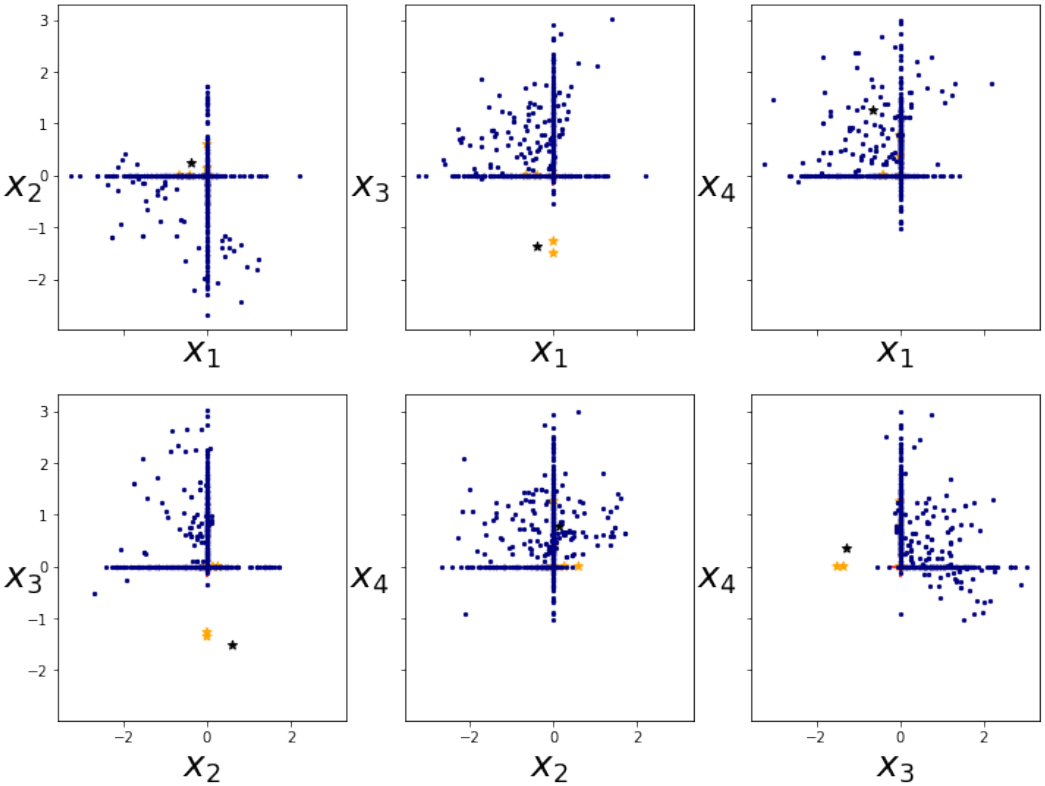

Figure 7: Initial points converging to $\mathbf{x}^{*(4)} = [-0.66, 0, 0, 1.26]^\top$ whose support is $\text{supp}(\mathbf{x}^{*(4)}) = \{1, 4\}$ and the function value is $f^{*(4)} = 0.25$ (frequency out of 2000 experiments is 554 and seed=152,315).

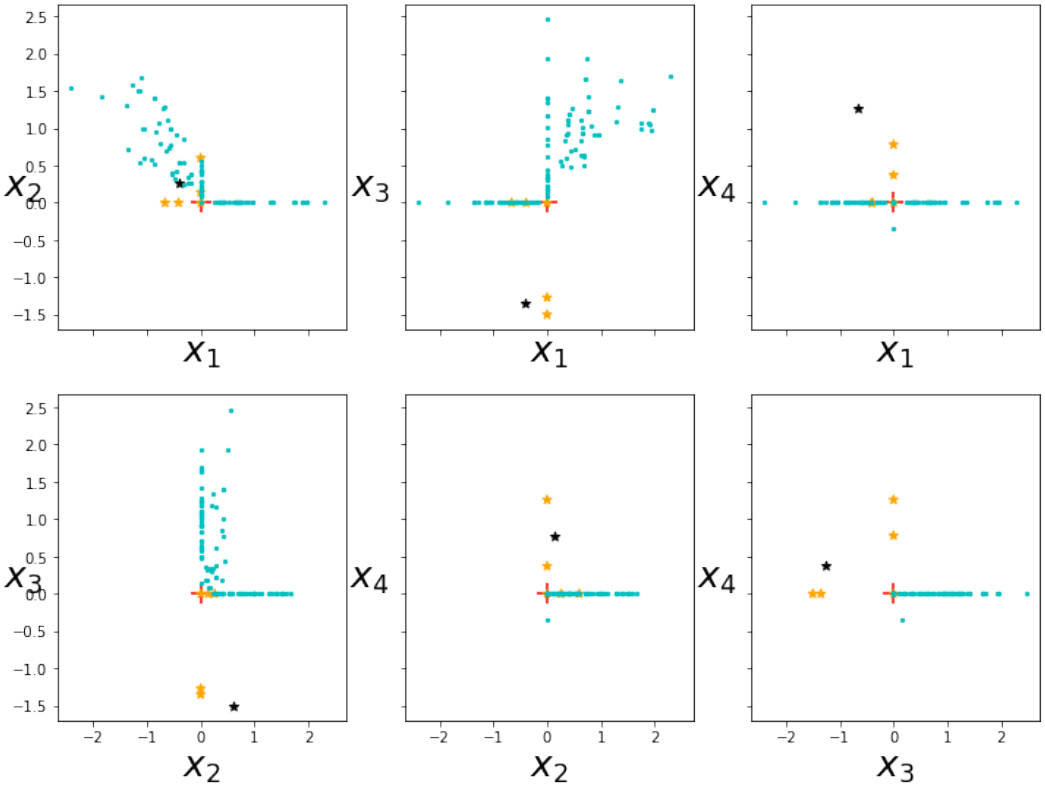

Figure 8: Initial points converging to $\mathbf{x}^{*(5)} = [-0.39, 0.26, 0, 0]^\top$ whose support is $\text{supp}(\mathbf{x}^{*(5)}) = \{1, 2\}$ and the function value is $f^{*(5)} = 0.53$ (frequency out of 2000 experiments is 116 and seed=152,315).

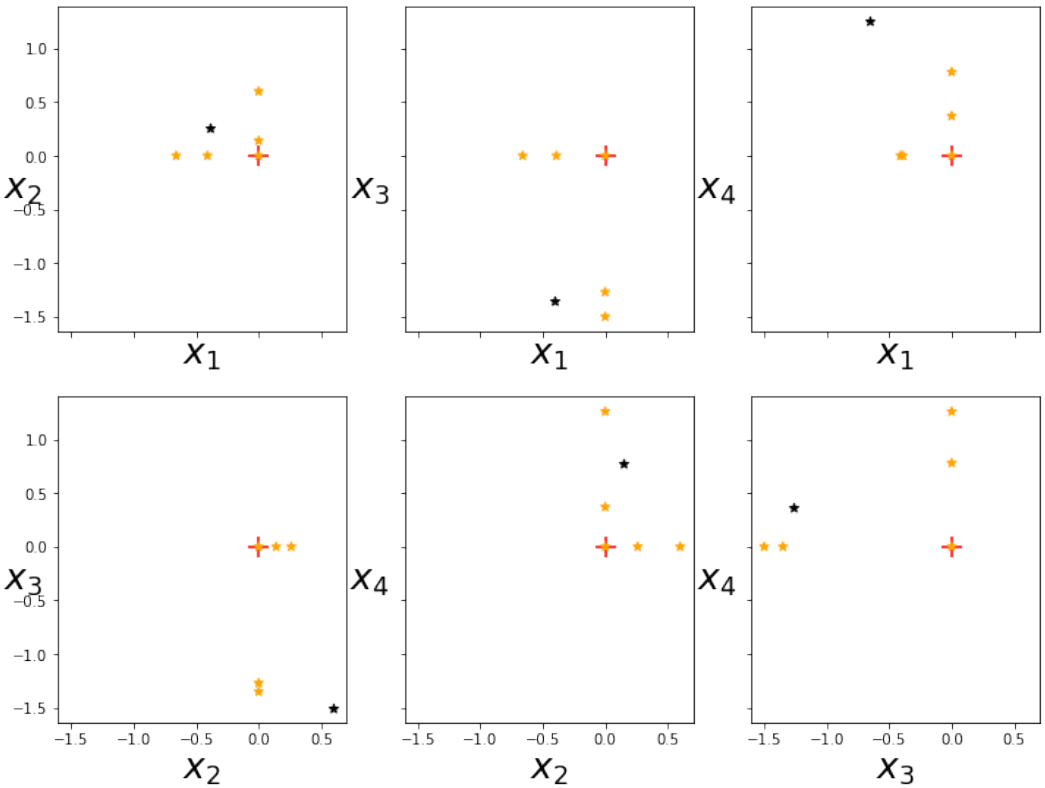

Figure 9: Initial points converging to $\mathbf{x}^{*(6)} = [0, 0.14, 0, 0.78]^\top$ whose support is $\text{supp}(\mathbf{x}^{*(6)}) = \{2, 4\}$ and the function value is $f^{*(6)} = 0.54$ (frequency out of 2000 experiments is 0 and seed=152,315).