# Sample Size Selection in Optimization Methods for Machine Learning

Xialiang Dou, Mingrui Zhang, Dongming Huang

22. December 2014

Setting:

$$N = 10000, p = 6, X \in R^{n \times p}$$

$$Y_i \in \{0, 1\}, P(Y_i = 1|X_i) = \frac{exp(w^T X_i)}{1 + exp(w^T X_i)}, i = 1, \cdots, N$$

$$J(w) = -\frac{1}{N} \sum_{i=1}^{N} log \frac{exp(w^T X_i)}{1 + exp(w^T X_i)}$$

**Algorithm 3.1: Dynamic Sample Gradient Algorithm**

Choose an initial iterate $w_0$, an initial sample $\mathcal{S}_0$, and a constant $\theta \in (0, 1)$.
Set $k \leftarrow 0$
**Repeat** until a convergence test is satisfied:

1 Compute $d_k = -\nabla J_{\mathcal{S}_k}(w_k)$
2 Line Search: compute steplength $\alpha_k > 0$ such that
$$J_{\mathcal{S}_k}(w_k + \alpha_k d_k) < J_{\mathcal{S}_k}(w_k)$$
3 Define a new iterate: $w_{k+1} = w_k + \alpha_k d_k$.
4 Set $k \leftarrow k + 1$.
5 Choose a sample $\mathcal{S}_k$ such that $|\mathcal{S}_k| = |\mathcal{S}_{k-1}|$.
6 Compute the sample variance defined in (3.6).
7 If condition (3.9) is not satisfied, augment $\mathcal{S}_k$ using formula (3.12).

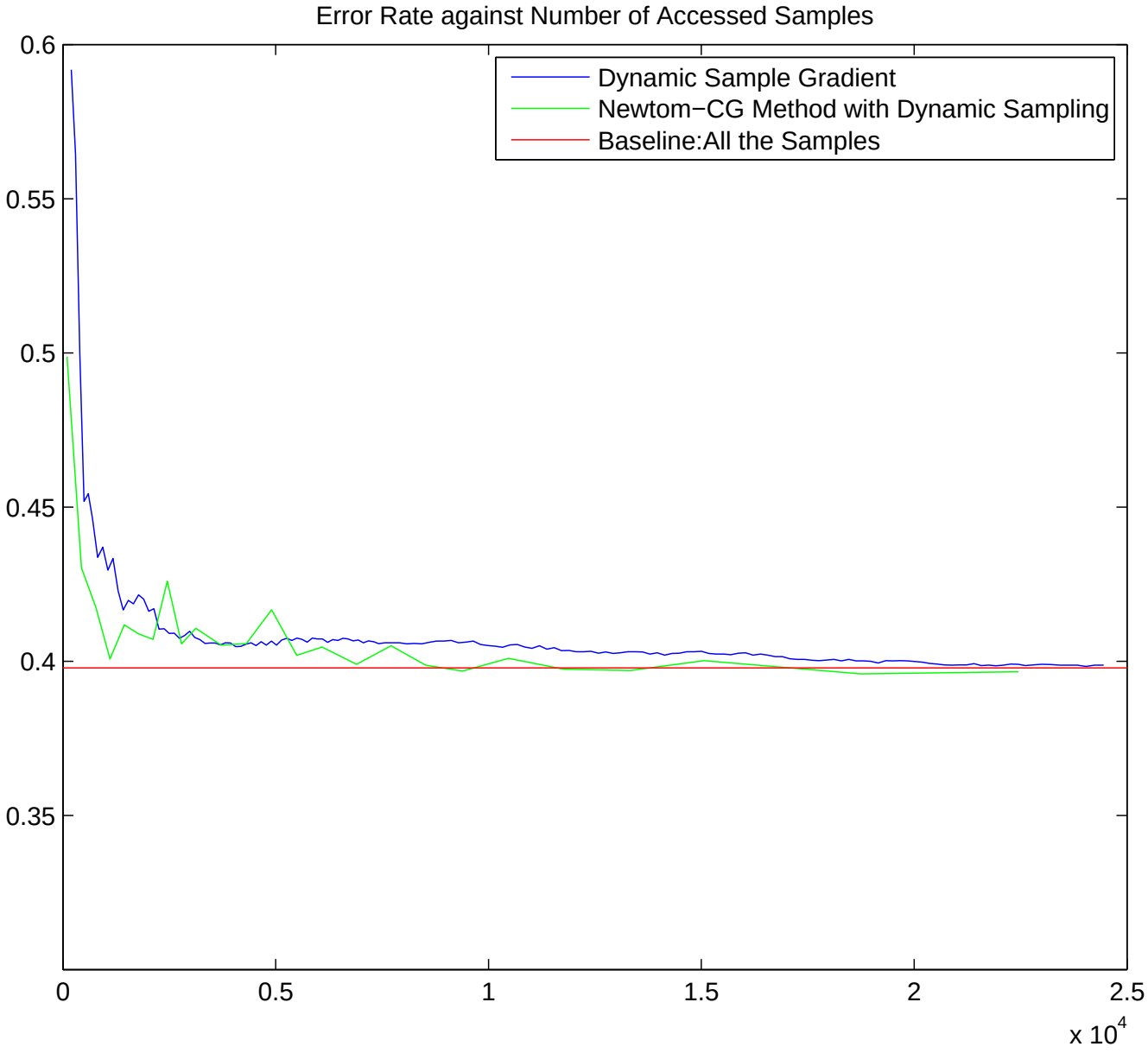
**Algorithm 5.2: Newton-CG Method with Dynamic Sampling**

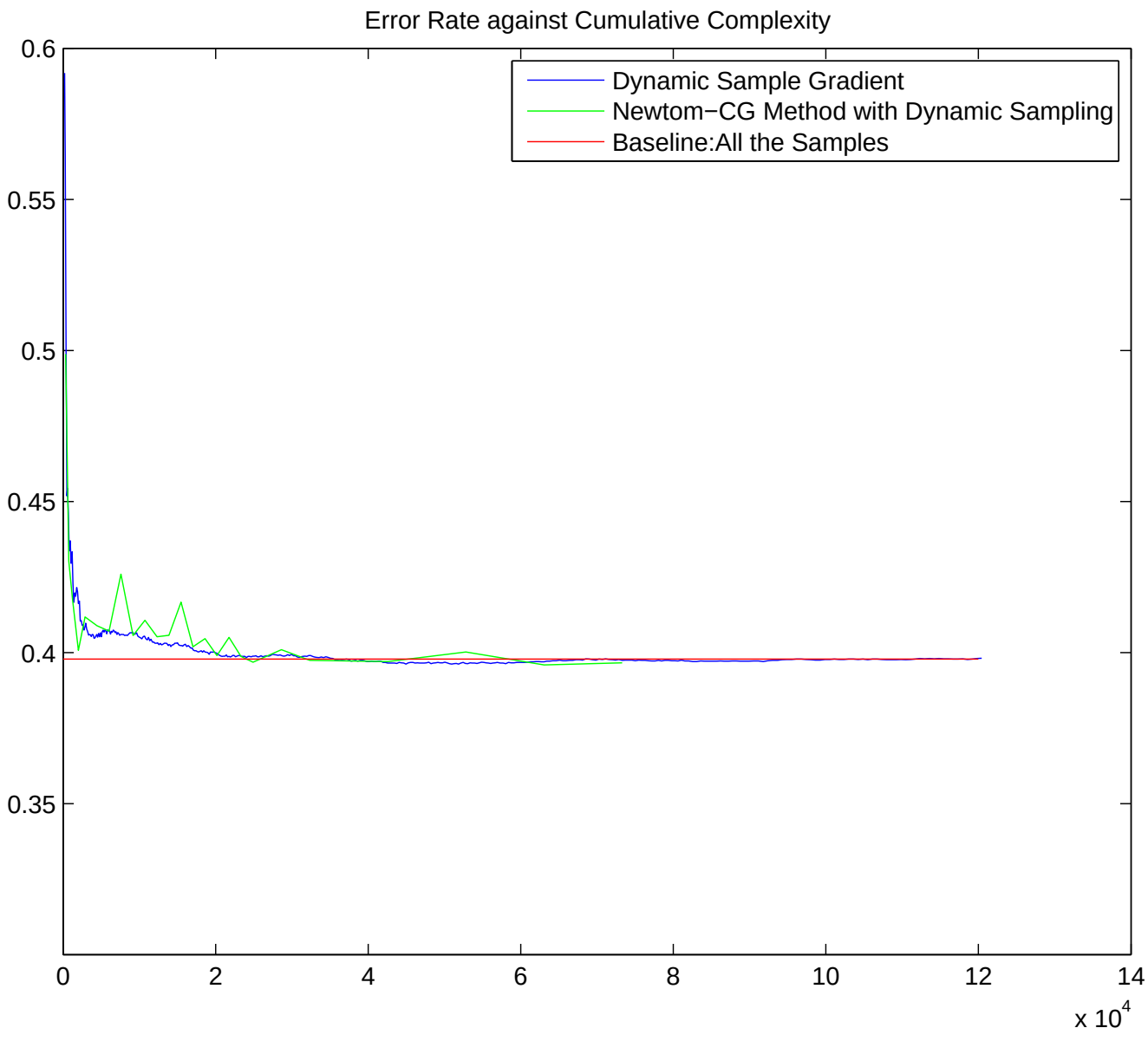
Initialize: Choose an initial iterate $w_0$, initial samples $\mathcal{H}_0 \subseteq \mathcal{S}_0$, and a sampling ratio $R$ such that $|\mathcal{H}_0| = R|\mathcal{S}_0|$. Choose constants $\theta \in (0, 1)$, $0 < c_1 < c_2 < 1$. Set $k \leftarrow 0$.

**Repeat** until a convergence test is satisfied:

1 Compute the search direction $d_k$ by means of Algorithm 5.1.
2 Compute a steplength $\alpha_k$ that satisfies the Wolfe conditions:
   1. $J_{\mathcal{S}_k}(w_k + \alpha_k d_k) \leq J_{\mathcal{S}_k}(w_k) + c_1 \alpha_k \nabla J_{\mathcal{S}_k}(w_k)^T d_k$
   2. $\nabla J_{\mathcal{S}_k}(w_k + \alpha_k d_k)^T d_k \geq c_2 \nabla J_{\mathcal{S}_k}(w_k)^T d_k$.
3 Define the new iterate: $w_{k+1} \leftarrow w_k + \alpha_k d_k$.
4 Increment the counter $k \leftarrow k + 1$.
5 Choose a sample $\mathcal{S}_k$ such that $|\mathcal{S}_k| = |\mathcal{S}_{k-1}|$.
6 If condition (3.9) is not satisfied, augment $\mathcal{S}_k$ using formula (3.12).
7 Select a sample $\mathcal{H}_k \subseteq \mathcal{S}_k$, such that $|\mathcal{H}_k| = \lceil R|\mathcal{S}_k| \rceil$.

- Parameter $\theta = 0.9$
- Initial sample size $S_0 = N \times 1\%$
- Hessian ratio $R = 0.2$


Error Rate against Number of Accessed Samples
Dynamic Sample Gradient
Newtom−CG Method with Dynamic Sampling
Baseline:All the Samples
0.6
0.55
0.5
0.45
0.4
0.35
0
0.5
1
1.5
2
2.5
x 10^4

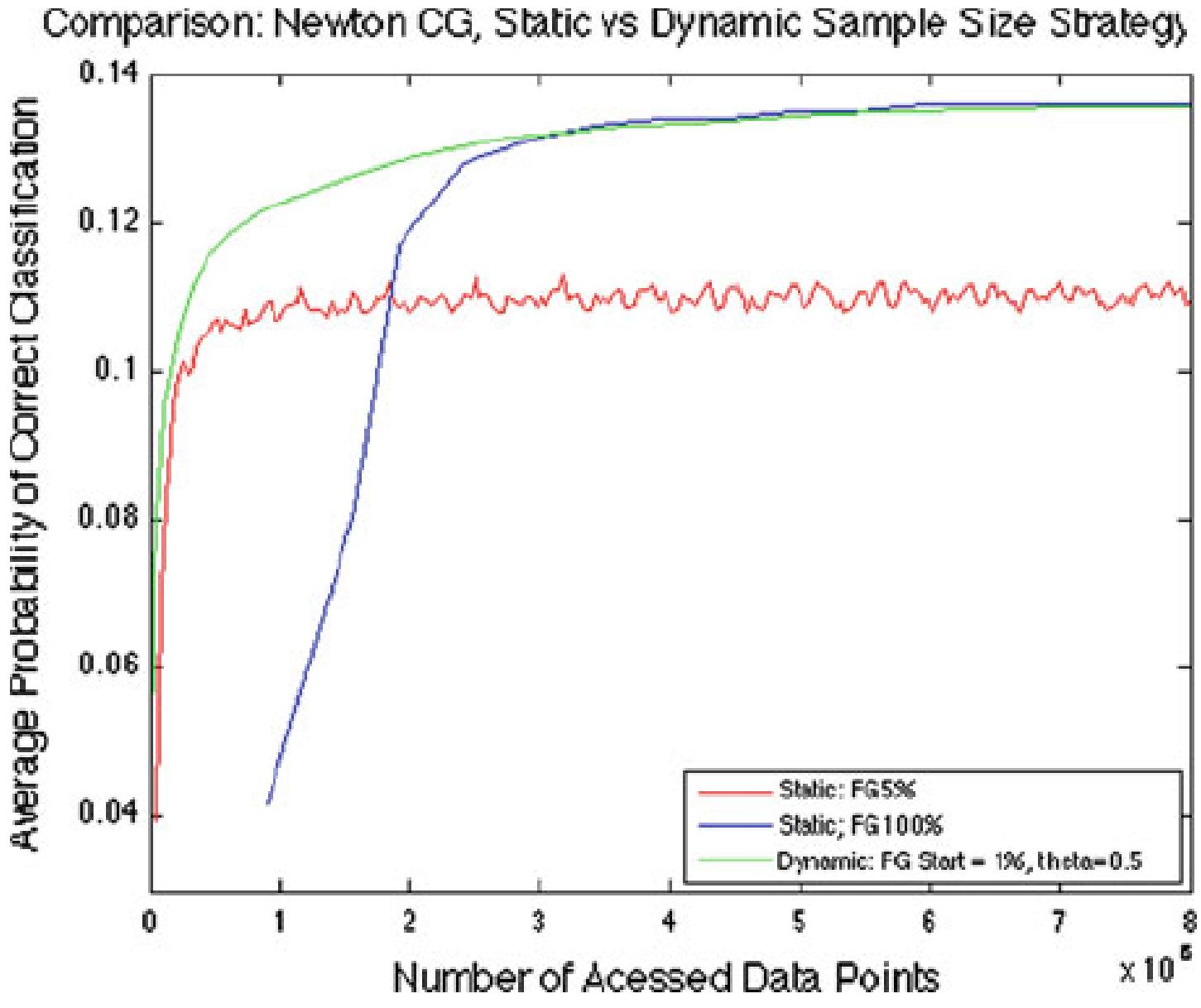

Error Rate against Cumulative Complexity
Dynamic Sample Gradient
Newtom−CG Method with Dynamic Sampling
Baseline:All the Samples
0.6
0.55
0.5
0.45
0.4
0.35
0
2
4
6
8
10
12
14
x 10^4

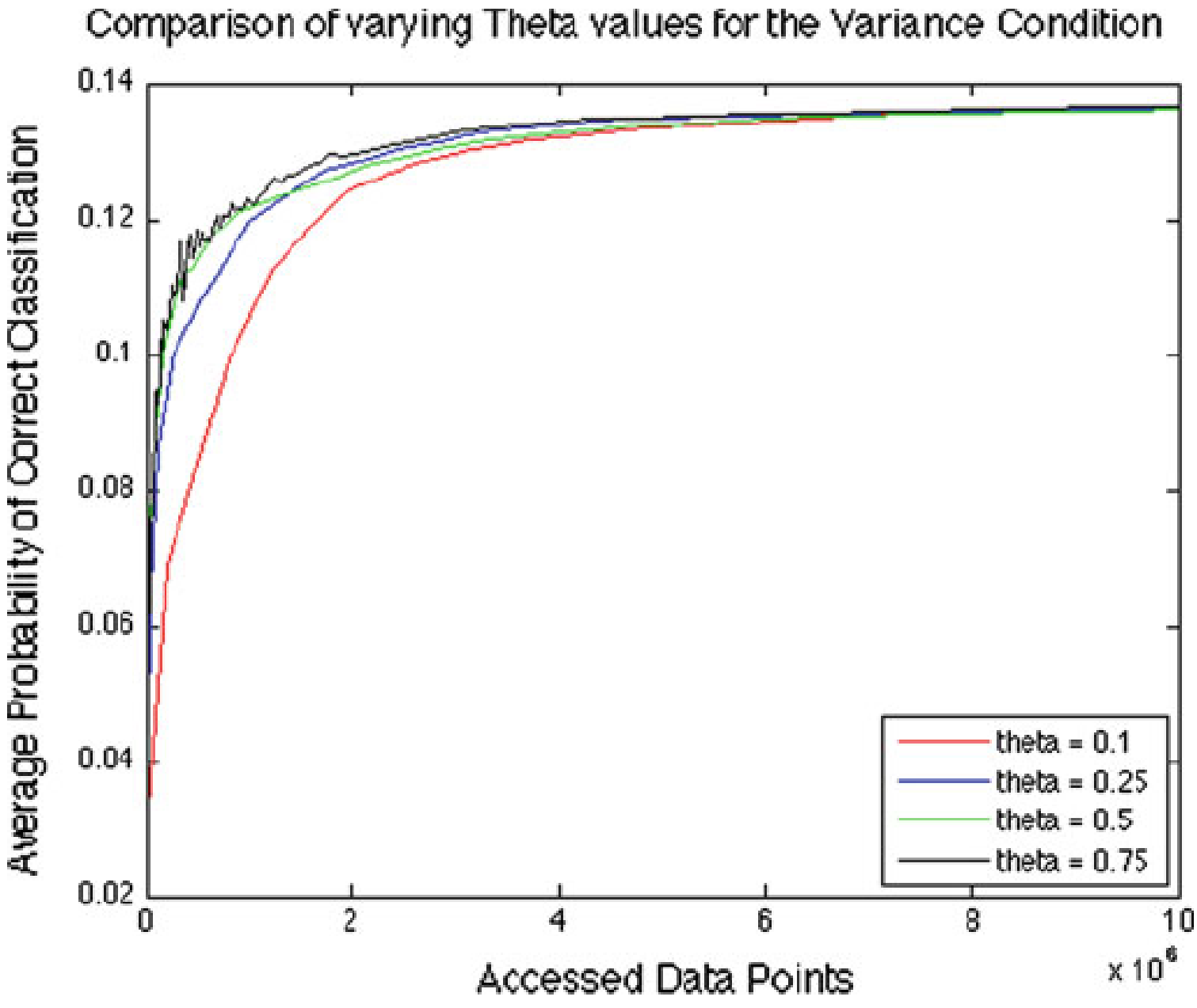
- Experiment on the paper,speech recognition problem from Google.
- $\|F\| = 79, \mathcal{C} = \{1, \cdots, 129\}, m = \|\mathcal{C}\| \times \|F\| = 10,191$
- N=168,776

$$
\begin{aligned}
A: &\quad \frac{\|\mathrm{Var}_{i\in\mathcal{S}}(\nabla\ell(w;i))\|_1}{|\mathcal{S}|}\\
B: &\quad \|\nabla J_{\mathcal{S}_k}(w)-\nabla J(w)\|_2^2\\
Y: &\quad \frac{\|\mathrm{Var}_{i\in\mathcal{H}_k}\left(\nabla^2\ell(w_k;i)\nabla J_{\mathcal{S}_k}(w_k)\right)\|_1}{|\mathcal{H}_k|\|\nabla J_{\mathcal{S}_k}(w_k)\|_2^2}\\
Z: &\quad \frac{\|[\nabla^2 J_{\mathcal{S}_k}(w_k)-\nabla^2 J_{\mathcal{H}_k}(w_k)]\nabla J_{\mathcal{S}_k}(w_k)\|_2^2}{\|\nabla J_{\mathcal{S}_k}(w_k)\|_2^2}.
\end{aligned}
$$

**Table 2** Analysis of error estimations

| Iter: $k$ | $A$ | $B$ | $Y$ | $Z$ |
|---|---|---|---|---|
| 1 | 0.041798576 | 0.039624485 | 0.015802753 | 0.016913783 |
| 2 | 0.040513236 | 0.03736721 | 0.033969807 | 0.034863612 |
| 3 | 0.036680293 | 0.035931856 | 0.027860283 | 0.019185297 |
| 4 | 0.028538358 | 0.028526197 | 0.017811839 | 0.017687138 |
| 5 | 0.02030112 | 0.01974306 | 0.015910543 | 0.014236308 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 13 | 0.001539071 | 0.001843071 | 0.002523053 | 0.002681272 |
| 14 | 0.000981444 | 0.001307763 | 0.0022572 | 0.002574446 |
| 15 | 0.000613751 | 0.000929579 | 0.000793887 | 0.001335829 |
| 16 | 0.000190385 | 0.00052025 | 0.000516926 | 0.00049049 |
| 17 | 0.000048608 | 0.000381851 | 0.00059497 | 0.0005979 |


Comparison: Newton CG, Static vs Dynamic Sample Size Strategy
Average Probability of Correct Classification
0.14
0.12
0.1
0.08
0.06
0.04
0
1
2
3
4
5
6
7
8
Number of Acessed Data Points
x 10^6
Static: FG5%
Static; FG100%
Dynamic: FG Start = 1%, theta=0.5


Comparison of varying Theta values for the Variance Condition
Average Probability of Correct Classification
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
2
4
6
8
10
x 10^6
Accessed Data Points
theta = 0.1
theta = 0.25
theta = 0.5
theta = 0.75