Attribute Efficient Linear Regression with Distribution-Dependent Sampling

Doron Kukliansky Ohad Shamir

Weizmann Institute of Science

ICML
July 2015

Motivation

Medical diagnosis classification task

- Problem: volunteers are willing to undergo only a small number of medical tests

Motivation

Medical diagnosis classification task

- Problem: volunteers are willing to undergo only a small number of medical tests
- Solution: limit the number of tests each volunteer undergoes, with the cost of using more volunteers

Problem Definition

What it is?

- Training set of m examples, each of dimension d
- At training time, can view only $k \ll d$ attributes
- At testing time, the model has access to all d attributes

Problem Definition

What it is?

- Training set of m examples, each of dimension d
- At training time, can view only $k \ll d$ attributes
- At testing time, the model has access to all d attributes

What it is not?

- Not missing data
- Not feature selection

Problem Definition

$m = 4,\ k = 4$

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \\ x_{41} & x_{42} & x_{43} & x_{44} \\ ? & ? & ? & ? \\ ? & ? & ? & ? \\ ? & ? & ? & ? \\ ? & ? & ? & ? \end{pmatrix}$$

$m = 8,\ k = 2$

$$\begin{pmatrix} x_{11} & ? & x_{13} & ? \\ x_{21} & ? & ? & x_{24} \\ ? & x_{32} & x_{33} & ? \\ x_{41} & ? & ? & x_{44} \\ ? & ? & x_{53} & x_{54} \\ ? & x_{62} & x_{63} & ? \\ x_{71} & ? & x_{73} & ? \\ ? & ? & x_{83} & x_{84} \end{pmatrix}$$

1. Can we learn using the same total number of attributes?

Problem Definition

$m = 4,\ k = 4$

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \\ x_{41} & x_{42} & x_{43} & x_{44} \\ ? & ? & ? & ? \\ ? & ? & ? & ? \\ ? & ? & ? & ? \\ ? & ? & ? & ? \end{pmatrix}$$

$m = 6,\ k = 2$

$$\begin{pmatrix} x_{11} & ? & x_{13} & ? \\ x_{21} & ? & ? & x_{24} \\ ? & x_{32} & x_{33} & ? \\ x_{41} & ? & ? & x_{44} \\ ? & ? & x_{53} & x_{54} \\ ? & x_{62} & x_{63} & ? \\ ? & ? & ? & ? \\ ? & ? & ? & ? \end{pmatrix}$$

2. Can we learn using the a smaller total number of attributes?

Our Focus

We Focus on:

- Stochastic i.i.d. data
- Linear regression: $\hat{y}_t = \langle \mathbf{w}, \mathbf{x}_t \rangle$
  - Under 2-norm constraint: $\|\mathbf{w}\|_2 \leq B$
- Squared loss: $\ell(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$

Our Focus

We Focus on:

- Stochastic i.i.d. data
- Linear regression: $\hat{y}_t = \langle \mathbf{w}, \mathbf{x}_t \rangle$
  - Under 2-norm constraint: $\|\mathbf{w}\|_2 \leq B$
- Squared loss: $\ell(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$

Goal:

- Minimize excess risk: $\mathbb{E}[L_{\mathcal{D}}(\mathbf{w}_{\text{output}})] - L_{\mathcal{D}}(\mathbf{w}^*)$

Linear Regression Under 2-norm Constraint

- Goal: loss minimization - optimization problem
- Solution: Stochastic Gradient Descent (SGD)
- Algorithm template [Cesa-Bianchi et al. (2011)]:
  1. Scan the training set
  2. For each example build an unbiased gradient estimator based on sampling k attributes
  3. Feed into SGD

Gradient Estimate

- Squared loss for $\mathbf{x}_t$:
$$\ell(\mathbf{w}; \mathbf{x}_t, y_t) = \frac{1}{2}\left(\langle \mathbf{w}, \mathbf{x}_t \rangle - y_t\right)^2$$
- Gradient for $\mathbf{x}_t$:
$$\nabla \ell(\mathbf{w}; \mathbf{x}_t, y_t) = \left(\langle \mathbf{w}, \mathbf{x}_t \rangle - y_t\right) \cdot \mathbf{x}_t$$

Gradient Estimate

- Squared loss for $\mathbf{x}_t$:
$$\ell(\mathbf{w}; \mathbf{x}_t, y_t) = \frac{1}{2}\left(\langle \mathbf{w}, \mathbf{x}_t \rangle - y_t\right)^2$$
- Gradient for $\mathbf{x}_t$:
$$\nabla \ell(\mathbf{w}; \mathbf{x}_t, y_t) = \left(\langle \mathbf{w}, \mathbf{x}_t \rangle - y_t\right) \cdot \mathbf{x}_t$$
- Estimation:
  - Sample $k-1$ attributes with probabilities q_i and set $\widetilde{\mathbf{x}}_t = \frac{1}{k-1}\sum_{r=1}^{k-1} \frac{1}{q_{i_{t,r}}} \mathbf{x_t}\left[i_{t,r}\right] \mathbf{e}_{i_{t,r}}$, or
$$\widetilde{\mathbf{x}_t} = \frac{1}{k-1}(0, ..., \frac{1}{q_{i_{t,1}}}\mathbf{x_t}\left[i_{t,1}\right], 0, ..., \frac{1}{q_{i_{t,k-1}}}\mathbf{x}_t\left[i_{t,k-1}\right], 0, ..., 0)$$
  - Sample 1 attribute with probability $p_{j_t} = w_{t,j_t}^2 / \left\|\mathbf{w}_t\right\|_2^2$ and set $\widetilde{\phi}_t = \frac{w_{t,j}}{p_{j_t}} \mathbf{x}_t\left[j_t\right] - y_t$
  - Set $\widetilde{\mathbf{g}}_t = \widetilde{\phi}_t \cdot \widetilde{\mathbf{x}}_t$

AERR Algorithm [Hazan and Koren (2012)]

- Use uniform sampling: $q_i = \frac{1}{d} \quad \forall i \in [d]$
- Expected risk bound: $\mathrm{O}\left(\sqrt{\frac{d}{km}}\right)$
  - Can learn using the same total number of attributes as the full-information scenario

AERR Algorithm [Hazan and Koren (2012)]

- Use uniform sampling: $q_i = \frac{1}{d} \quad \forall i \in [d]$
- Expected risk bound: $\mathrm{O}\left(\sqrt{\frac{d}{km}}\right)$
  - Can learn using the same total number of attributes as the full-information scenario

Can we do better?

AERR Algorithm [Hazan and Koren (2012)]

- Use uniform sampling: $q_i = \frac{1}{d} \quad \forall i \in [d]$
- Expected risk bound: $\mathrm{O}\left(\sqrt{\frac{d}{km}}\right)$
  - Can learn using the same total number of attributes as the full-information scenario

Can we do better?

No.

AERR Algorithm [Hazan and Koren (2012)]

- Use uniform sampling: $q_i = \frac{1}{d} \quad \forall i \in [d]$
- Expected risk bound: $O\left(\sqrt{\frac{d}{km}}\right)$
  - Can learn using the same total number of attributes as the full-information scenario

Can we do better?

No.

- They also prove a corresponding lower bound

Algorithm Template Analysis

- How should we pick the q_i-s?

Algorithm Template Analysis

- How should we pick the q_i-s?
- Can show that the risk bound depends on

$$\mathbb{E}\left[\left\|\widetilde{\mathbf{g}}_t\right\|_2^2\right] \leq 4B^2\left(\frac{1}{k-1}\sum_{i=1}^{d}\frac{1}{q_i}\mathbb{E}\left[x_i^2\right]+1\right)$$

Algorithm Template Analysis

- How should we pick the q_i-s?
- Can show that the risk bound depends on

$$\mathbb{E}\left[\left\|\widetilde{\mathbf{g}}_t\right\|_2^2\right] \leq 4B^2\left(\frac{1}{k-1}\sum_{i=1}^{d}\frac{1}{q_i}\mathbb{E}\left[x_i^2\right]+1\right)$$

- Try to minimize as a function of the q_i-s

Algorithm Template Analysis

- How should we pick the q_i-s?
- Can show that the risk bound depends on

$$\mathbb{E}\left[\left\|\widetilde{\mathbf{g}}_t\right\|_2^2\right] \leq 4B^2\left(\frac{1}{k-1}\sum_{i=1}^{d}\frac{1}{q_i}\mathbb{E}\left[x_i^2\right]+1\right)$$

- Try to minimize as a function of the q_i-s
- Solution:

$$q_i = \frac{\sqrt{\mathbb{E}\left[x_i^2\right]}}{\sum_{j=1}^{d}\sqrt{\mathbb{E}\left[x_j^2\right]}}.$$

Known Second Moments - DDAERR

- The optimal sampling probabilities can be calculated if second moments of attributes are known
- This is the DDAERR (Distribution-Dependent Attribute Efficient Ridge Regression) algorithm
  - Risk bound:

$$O\left(\sqrt{\frac{\left\|\mathbb{E}\left[\mathbf{x}^2\right]\right\|_{\frac{1}{2}}+k}{km}}\right)$$

DDAERR Results

- By definition $\left\|\mathbb{E}\left[\mathbf{x}^2\right]\right\|_{\frac{1}{2}} = \left(\sum_{i=1}^{d}\sqrt{\mathbb{E}\left[\mathbf{x}_i^2\right]}\right)^2$
- Assuming $\|\mathbf{x}\|_2 \leq 1$:
  - $\left\|\mathbb{E}\left[\mathbf{x}^2\right]\right\|_{\frac{1}{2}} \leq d$
  - $\left\|\mathbb{E}\left[\mathbf{x}^2\right]\right\|_{\frac{1}{2}} = d$ only when all second moments are equal
  - On real-world data sets, typically $\left\|\mathbb{E}\left[\mathbf{x}^2\right]\right\|_{\frac{1}{2}} < d$
- The example in the lower bound satisfies $\left\|\mathbb{E}\left[\mathbf{x}^2\right]\right\|_{\frac{1}{2}} = d$

DDAERR Results

- By definition $\left\|\mathbb{E}\left[\mathbf{x}^2\right]\right\|_{\frac{1}{2}} = \left(\sum_{i=1}^{d}\sqrt{\mathbb{E}\left[\mathbf{x}_i^2\right]}\right)^2$
- Assuming $\|\mathbf{x}\|_2 \leq 1$:
  - $\left\|\mathbb{E}\left[\mathbf{x}^2\right]\right\|_{\frac{1}{2}} \leq d$
  - $\left\|\mathbb{E}\left[\mathbf{x}^2\right]\right\|_{\frac{1}{2}} = d$ only when all second moments are equal
  - On real-world data sets, typically $\left\|\mathbb{E}\left[\mathbf{x}^2\right]\right\|_{\frac{1}{2}} < d$
- The example in the lower bound satisfies $\left\|\mathbb{E}\left[\mathbf{x}^2\right]\right\|_{\frac{1}{2}} = d$

Conclusion

The DDAERR algorithm uses a smaller total amount of attributes compared to the full-information algorithms on real-world data sets.

Unknown Second Moments - Two Phased DDAERR

- What if we don’t know second moments?

Unknown Second Moments - Two Phased DDAERR

- What if we don’t know second moments?
- Solution - two phased method:
  1. Estimate second moments
  2. Use UCB-style moment estimates

Unknown Second Moments - Two Phased DDAERR

- What if we don't know second moments?
- Solution - two phased method:
  1. Estimate second moments
  2. Use UCB-style moment estimates
- Estimation phase samples attributes uniformly
  - Run the AERR algorithm $\Rightarrow$ no data is wasted

1-norm Constraint

- Again linear regression and squared loss, but this time with a different norm constraint: $\|\mathbf{w}\|_1 \leq B$

1-norm Constraint

- Again linear regression and squared loss, but this time with a different norm constraint: $\|\mathbf{w}\|_1 \leq B$

- Solution similar to the 2-norm constraint:
  - Based on the Exponentiated Gradient algorithm instead of the Stochastic Gradient Descent algorithm
  - Uses different sampling distributions:

$$q_i = \frac{\mathbb{E}\left[x_i^2\right]}{\sum_{j=1}^{d} \mathbb{E}\left[x_j^2\right]}$$

  - Risk bound:

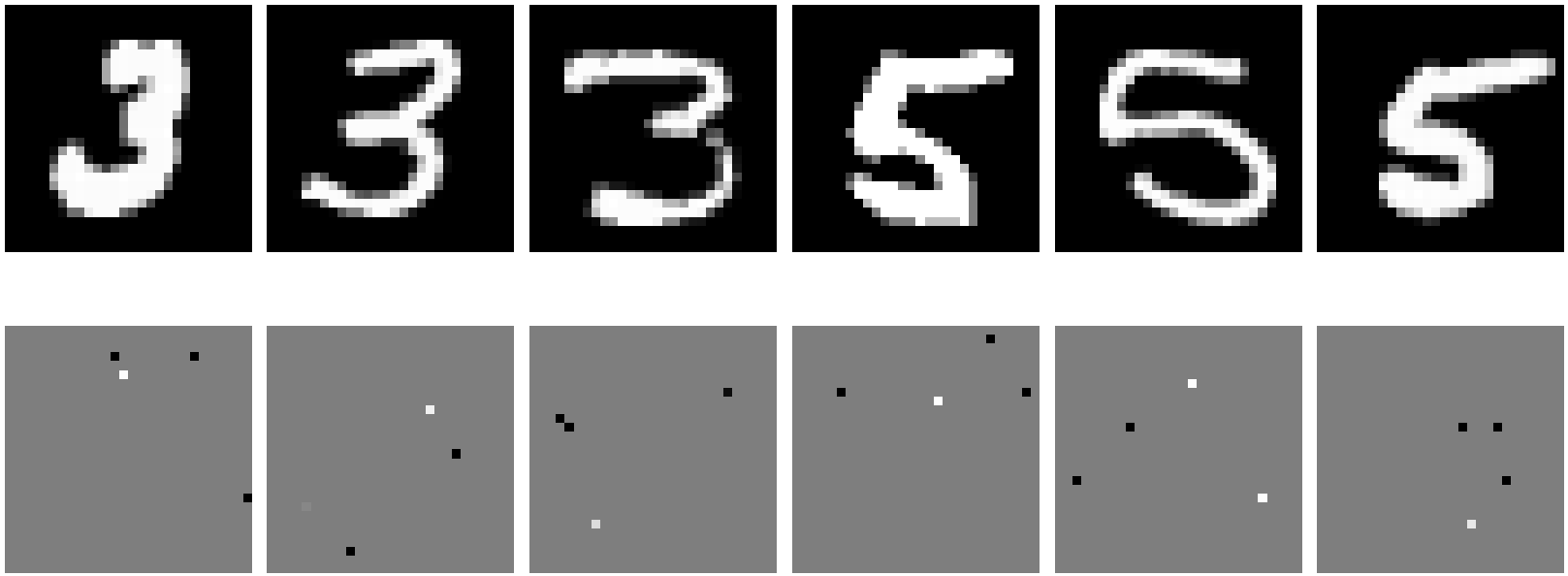
$$\mathrm{O}\left(\sqrt{\frac{\left(\left\|\mathbb{E}\left[\mathbf{x}^2\right]\right\|_1 + k\right)\log d}{km}}\right)$$

Experiment - MNIST Data Set

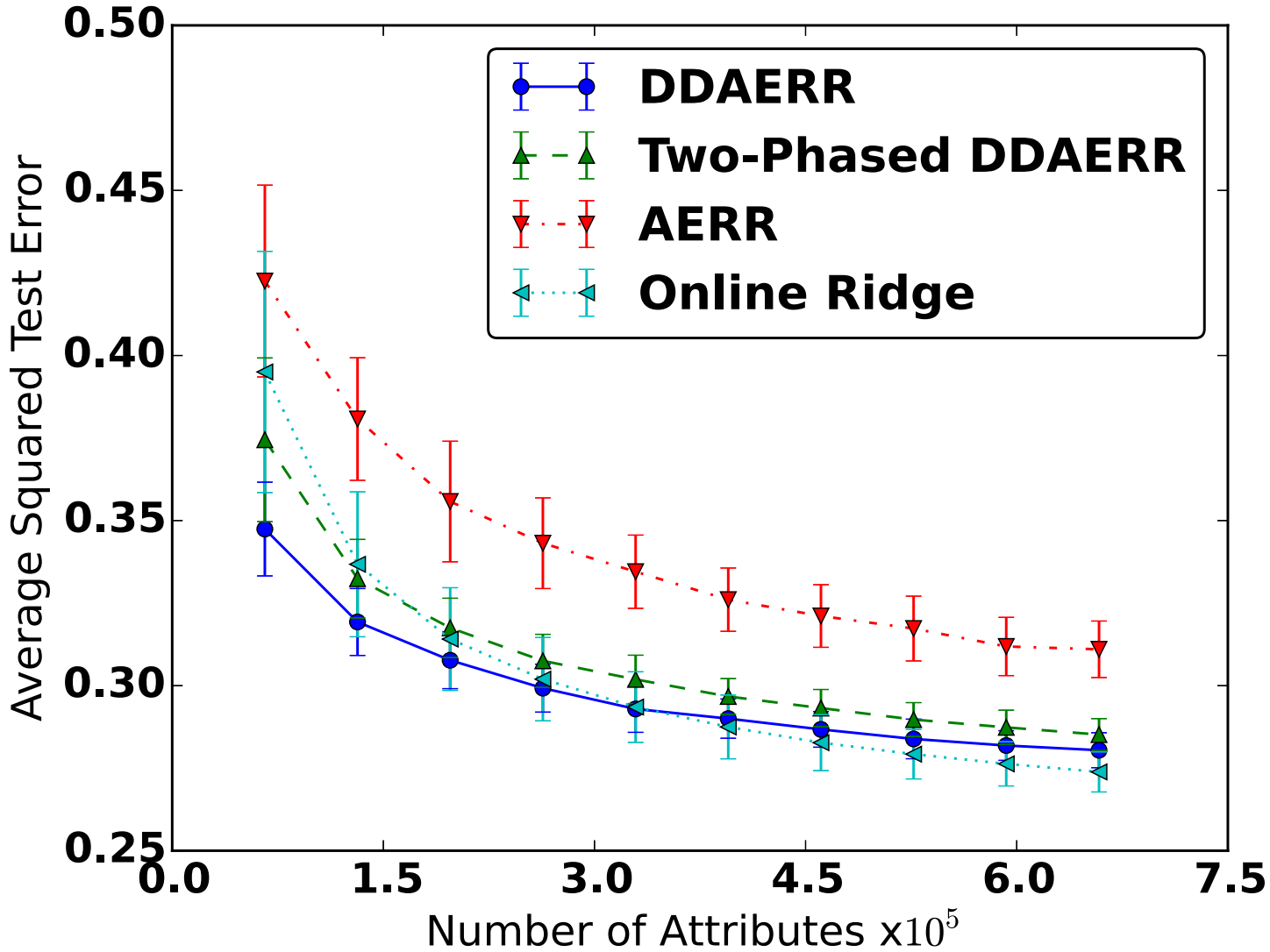
- Consist of gray scale hand-written digits
- Standard machine learning data set
- We focused on a classification problem: ’3’ vs. ’5’

Experiment - MNIST Data Set

- Consist of gray scale hand-written digits
- Standard machine learning data set
- We focused on a classification problem: ’3’ vs. ’5’

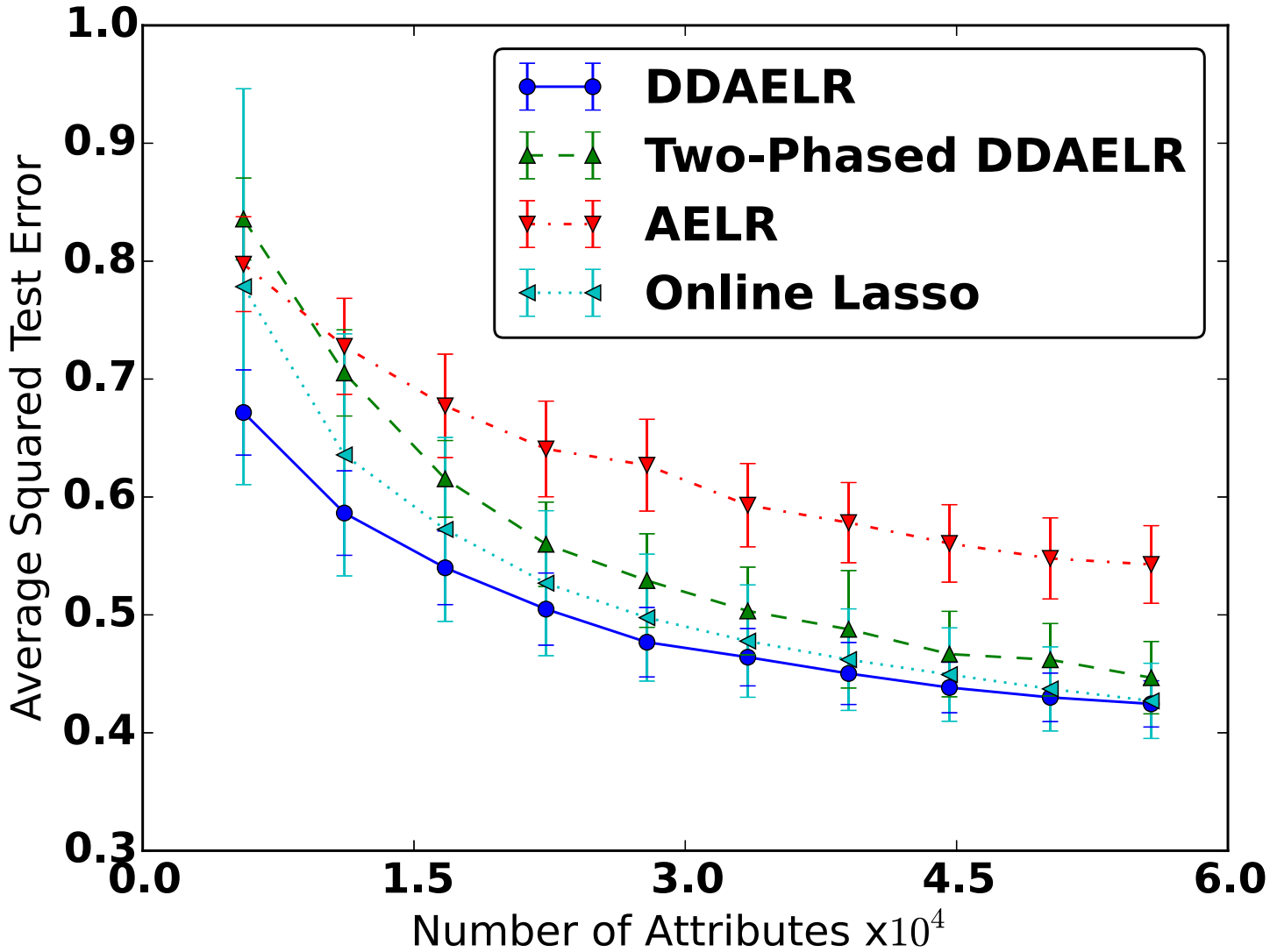


MNIST Data Set - 2-norm Constraint



Test error for the algorithms with $k = 56$ in the 2-norm scenario over the classification task ”3” vs. ”5” in the MNIST data set.

MNIST Data Set - 1-norm Constraint



Test error for the algorithms with $k = 4$ in the 1-norm scenario over the classification task ”3” vs. ”5” in the MNIST data set.

Summary

- Distribution-dependent sampling and bound
- Uses less attributes than full-information algorithms - useful also in general budgeted learning settings
- ”Broken” lower bound for real world data sets
- Method can potentially be effective in other partial-information learning scenarios

Summary

- Distribution-dependent sampling and bound
- Uses less attributes than full-information algorithms - useful also in general budgeted learning settings
- ”Broken” lower bound for real world data sets
- Method can potentially be effective in other partial-information learning scenarios

Thanks!