Dual Averaging Method for Online Graph-Structured Sparsity

Baojian Zhou[1,2], Feng Chen[1], and Yiming Ying[2]

[1]Department of Computer Science, [2]Department of Mathematics and Statistics, University at Albany, NY, USA

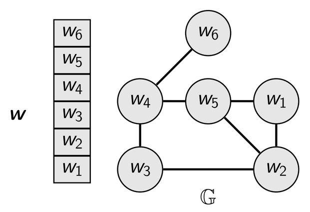
UNIVERSITY AT ALBANY
State University of New York

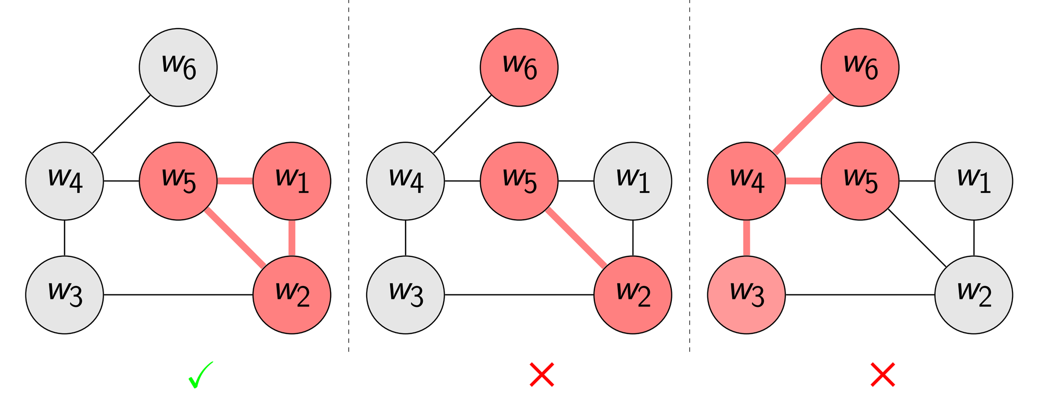
Online Graph-structured Learning

We study an online graph-structured learning problem, which is to minimize the regret as defined in the following

$$R(T,\mathcal{M}(\mathbb{M})) := \sum_{t=1}^{T} f_t(w_t,\{x_t,y_t\}) - \min_{w\in\mathcal{M}(\mathbb{M})} \sum_{t=1}^{T} f_t(w,\{x_t,y_t\}),$$

- $f_t(w,\{x_t,y_t\})$ is a convex loss
- $\mathcal{M}(\mathbb{M})$ models structured sparsity such as connected subgraphs, dense subgraphs, and subgraphs isomophic to a query graph, . . .





Weighted Graph Model $\mathbb{M} = \{S : |S| \le 3, S \text{ is connected}\}$ Hegde et al. (2015a).

Main Idea

An intuitive way to do this is to use online projected gradient descent Zinkevich (2003) where the algorithm needs to solve the following projection at iteration t:

$$w_{t+1} = \mathrm{P}(w_t - \eta_t \nabla f_t(w_t), \mathcal{M}(\mathbb{M})), \quad (1)$$

where η_t is the learning rate and P is the projection operator onto $\mathcal{M}(\mathbb{M})$, i.e., $\mathrm{P}(\cdot,\mathcal{M}(\mathbb{M})) : \mathbb{R}^p \to \mathbb{R}^p$ is defined as

$$\mathrm{P}(w,\mathcal{M}(\mathbb{M})) = \arg\min_{x\in\mathcal{M}(\mathbb{M})} \|w - x\|^2. \quad (2)$$

However, there are two essential drawbacks of online PGD

- The projection in (1) only uses single gradient $\nabla f_t(w_t)$ which is too noisy (large variance) to capture the graph-structured information at each iteration;
- The training samples coming later are less important than these coming earlier due to the decay of learning rate η_t.

Inspired by dual averaging-based methods, at each iteration, our method updates w_t by using the following minimization step:

$$w_{t+1} = \arg\min_{w\in\mathcal{M}(\mathbb{M})} \left\{ \left\langle \frac{1}{t+1}\sum_{i=0}^{t} g_i, w \right\rangle + \frac{\beta_t}{2t}\|w\|_2^2 \right\}, \quad (3)$$

where β_t is to control the learning rate implicitly and g_i is a subgradient in $\partial f_i(w,\{x_i,y_i\}) = \{g : f_i(z,\{x_i,y_i\}) \ge f_i(w,\{x_i,y_i\}) + \langle g, z - w\rangle, \forall z \in \mathcal{M}(\mathbb{R})\}$. The minimization step (3) has the following equivalent projection problems, specified in the following Theorem.

Theorem: Assume $\beta_t = \gamma\sqrt{t}$, where $\gamma > 0$ and denote $\bar{s}_{t+1} = \frac{1}{t+1}\sum_{i=0}^{t} g_i$. The minimization step of (3) can be expressed as the following two equivalent optimization problems:

$$\max_{S\in\mathbb{M}} \|P(-\frac{\sqrt{t}\bar{s}_{t+1}}{\gamma}, S)\|_2^2 \quad (4)$$

$$\min_{S\in\mathbb{M}} \|-\frac{\sqrt{t}\bar{s}_{t+1}}{\gamma} - P(-\frac{\sqrt{t}\bar{s}_{t+1}}{\gamma}, S)\|_2^2, \quad (5)$$

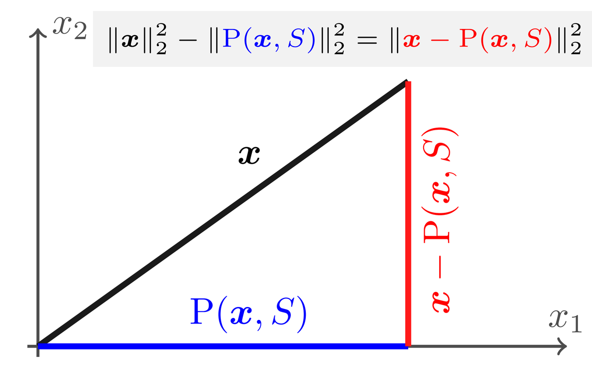
where $P(s,S)$ is the projection operator that projects s onto the subspace spanned by S.

The original minimization problem can be equivalently expressed as

$$w_{t+1} = \arg\min_{w\in\mathcal{M}(\mathbb{M})} \left\{ \langle \bar{s}_{t+1}, w\rangle + \frac{\gamma}{2\sqrt{t}}\|w\|_2^2 \right\}$$
$$= \arg\min_{w\in\mathcal{M}(\mathbb{M})} \left\| w - \left(-\frac{\sqrt{t}}{\gamma}\bar{s}_{t+1}\right)\right\|_2^2,$$

Each step is essentially a projection !

Theorem Insight



Define $x := -\sqrt{t}\bar{s}_{t+1}$ and adding minimization to both sides

$$\min_{S\in\mathbb{M}} \left\{ \|x\|^2 - \|\mathrm{P}(x,S)\|^2 \right\} = \min_{S\in\mathbb{M}} \|x - \mathrm{P}(x,S)\|^2.$$

By moving the minimization into the negative term, we obtain

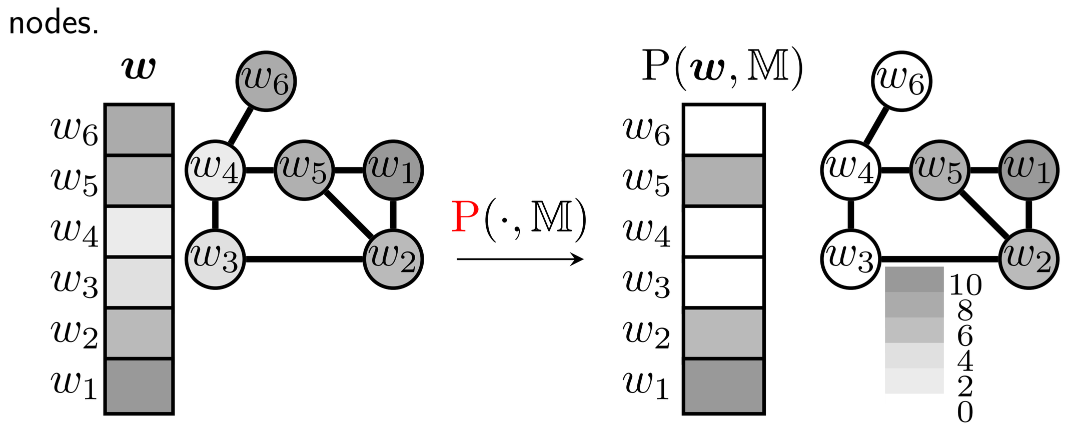
$$\|x\|^2 + \max_{S\in\mathbb{M}} \|\mathrm{P}(x,S)\|^2 = \min_{S\in\mathbb{M}} \|x - \mathrm{P}(x,S)\|^2.$$

Proposed Algorithms

Algorithm 1 GraphDA: Online Graph Dual Averaging Algorithm

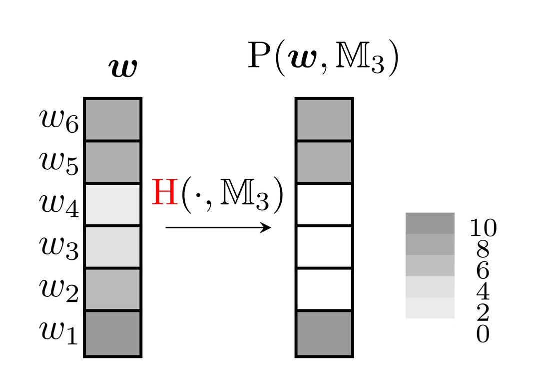
Input: γ, $\mathbb{M}$
$\bar{s}_0 = \mathbf{0}$, $w_0 = \mathbf{0}$
for $t = 0, 1, 2, \dots$ **do**
  receive $\{x_t, y_t\}$ and compute $g_t = \nabla f_t(w_t,\{x_t,y_t\})$
  $\bar{s}_{t+1} = \bar{s}_t + g_t$
  $b_{t+1} = \mathrm{P}(\bar{s}_{t+1}, \mathbb{M})$
  $w_{t+1} = \mathrm{P}(-\frac{\sqrt{t}}{\gamma} b_{t+1}, \mathbb{M})$
end for

Let $\mathbb{M} = \{S : |S| \le 3, S \text{ is connected }\}$. Finding a connected subgraph up to 3 nodes.



Graph Projection Operator Hegde et al. (2015b)

What if the graph information is not available ?



Sparse Projection Operator

Algorithm 2 DAIHT

Input: γ, $\mathbb{M}$
$\bar{s}_0 = \mathbf{0}$, $w_0 = \mathbf{0}$
for $t = 0, 1, 2, \dots$ **do**
  receive $\{x_t, y_t\}$ and compute $g_t = \nabla f_t(w_t,\{x_t,y_t\})$
  $\bar{s}_{t+1} = \bar{s}_t + g_t$
  $b_{t+1} = \mathrm{H}(\bar{s}_{t+1}, \mathbb{M})$
  $w_{t+1} = \mathrm{H}(-\frac{\sqrt{t}}{\gamma} b_{t+1}, \mathbb{M})$
end for

Time Complexity and Regret

The time complexity of GraphDA mainly depends on two projections. If we use the weighted-graph model, the per-iteration time cost could be

- non-sparse graph: $\mathcal{O}(p + |\mathbb{E}|\log^3(p))$ **Edge-dependent**
- sparse graph: $\mathcal{O}(p + p\log^3(p))$ **Nearly-linear !**

If $\mathcal{M}(\mathbb{M})$ is a **convex set**, then

- The regret can be bounded as: $R(T,\mathcal{M}(\mathbb{M})) = C \cdot \mathcal{O}(\sqrt{T})$, where C is a constant.
- If we assume further that the loss is strongly convex, then $\|w_T - w\|_2^2 = \mathcal{O}(\frac{\ln T}{T})$.

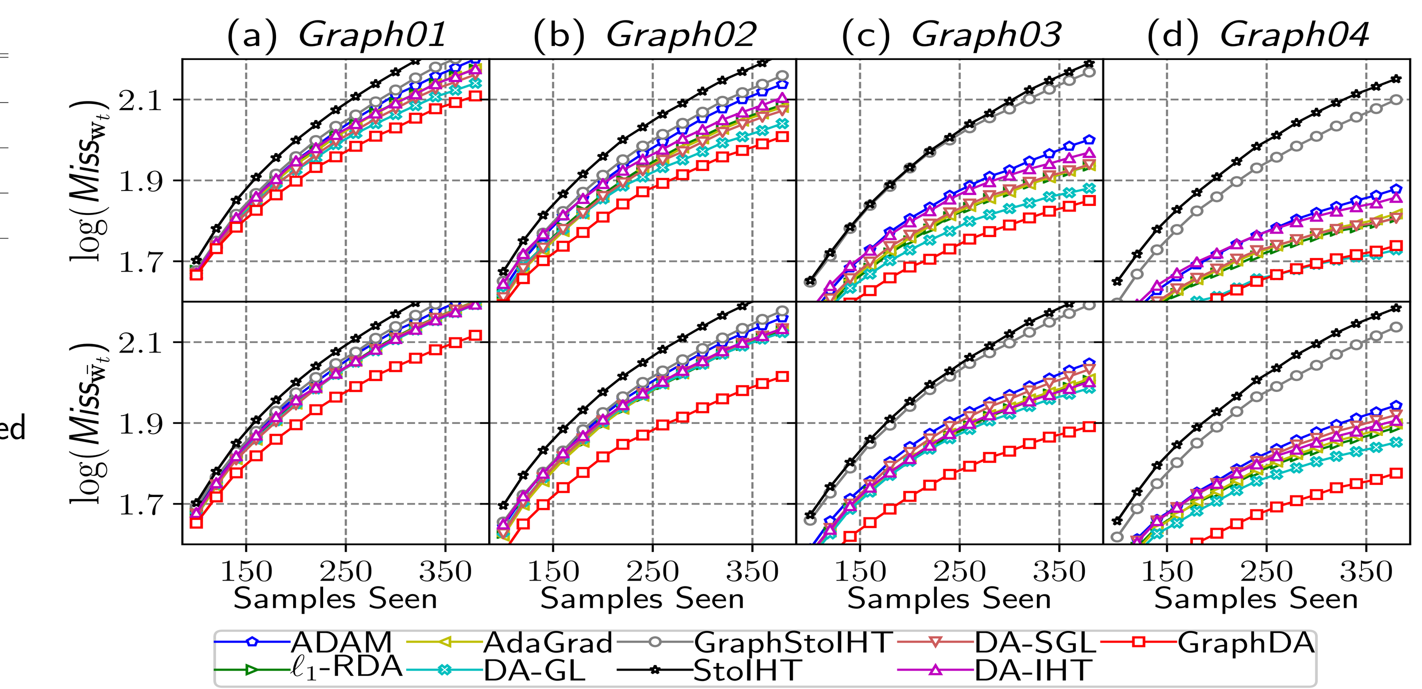
However, $\mathcal{M}(\mathbb{M})$ is not convex in our case. We leave the regret bound analysis of this case as a future work.

Experiments

Method	Proposed in
Adam	Kingma and Ba (2014)
ℓ_1-RDA	Xiao (2010)
DA-GL	Yang et al. (2010)
DA-SGL	Yang et al. (2010)
AdaGrad	Duchi et al. (2011)
StoIHT	Nguyen et al. (2017)
GraphStoIHT	Zhou et al. (2019)
DA-IHT	**This paper**
GraphDA	**This paper**

Dataset	$\lvert\mathbb{V}\rvert$	$\lvert\mathbb{E}\rvert$
Benchmark	1,089	2,112
MNIST	786	1,516
KEGG	5,372	78,545

non-sparse
sparse: convex-based
sparse: nonconvex-based



Aim to answer the following two questions:

- Can GraphDA achieve better classification performance?
- Can GraphDA learn stronger interpretative model through capturing more meaningful graph-structured features?

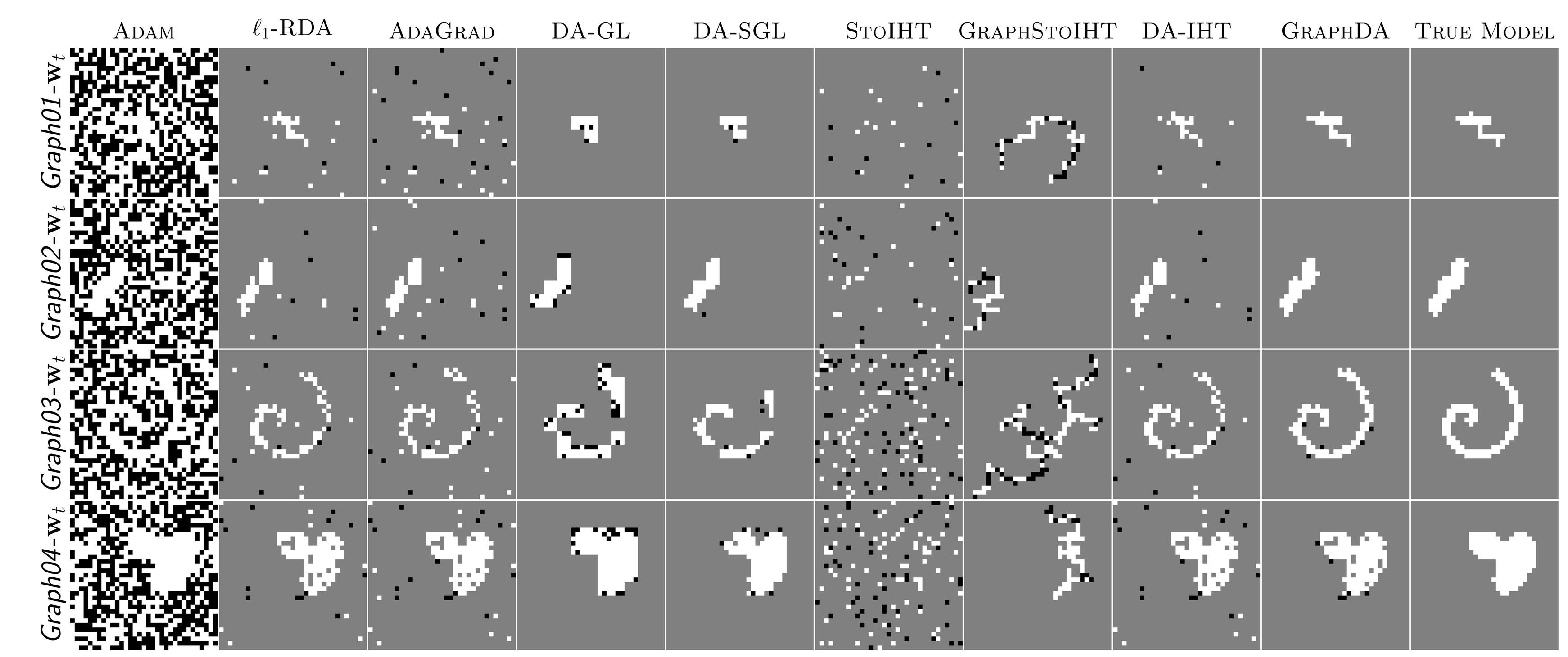
Evaluation Metric

- Area Under the ROC Curve(AUC) score
- Classification Accurracy(Acc)
- Nonzero Ratio

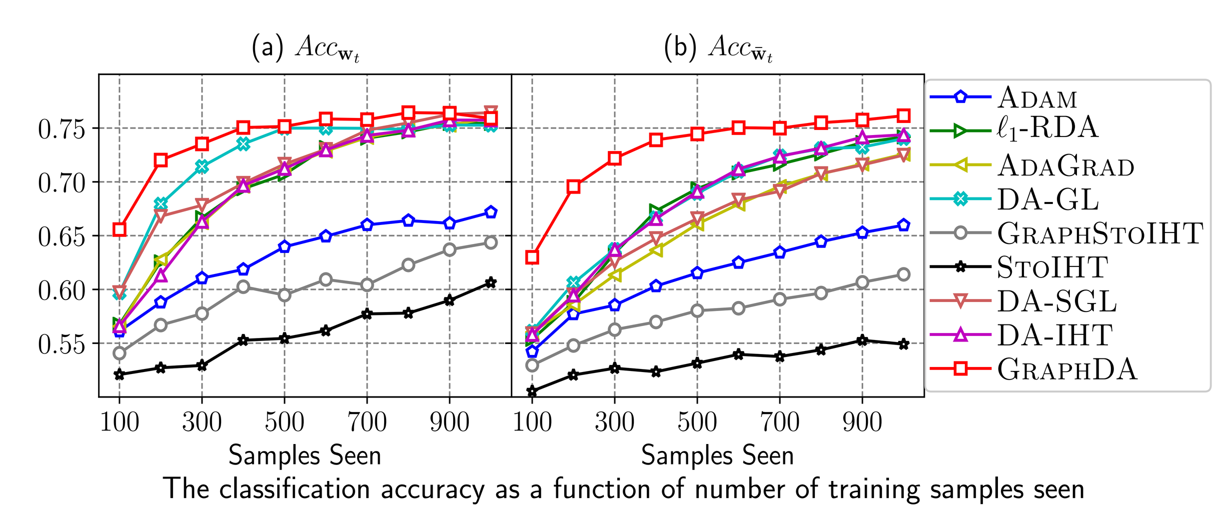
$$\mathrm{NR}_w = \frac{|\mathrm{supp}(w)|}{p}$$

Method	$AUC_{w_t,\bar{w}_t}$	$Acc_{w_t,\bar{w}_t}$	$NR_{w_t,\bar{w}_t}$
Adam	(0.618, 0.603)	(0.619, 0.603)	(100.0%, 100.0%)
ℓ_1-RDA	(0.693, 0.672)	(0.694, 0.673)	(11.58%, 83.60%)
AdaGrad	(0.696, 0.636)	(0.696, 0.637)	(11.33%, 100.0%)
DA-GL	(0.735, 0.666)	(0.735, 0.667)	(15.99%, 100.0%)
DA-SGL	(0.699, 0.647)	(0.699, 0.647)	(25.54%, 100.0%)
StoIHT	(0.552, 0.523)	(0.553, 0.523)	(7.79%, 40.62%)
GraphStoIHT	(0.603, 0.570)	(0.602, 0.570)	(7.84%, **22.06%**)
DA-IHT	(0.697, 0.666)	(0.697, 0.666)	(4.35%, 39.50%)
GraphDA	(**0.749**, **0.739**)	(**0.749**, **0.739**)	(**2.56%**, 32.12%)

- GraphDA has better classification performance.
- GraphDA has stronger model interpretability.

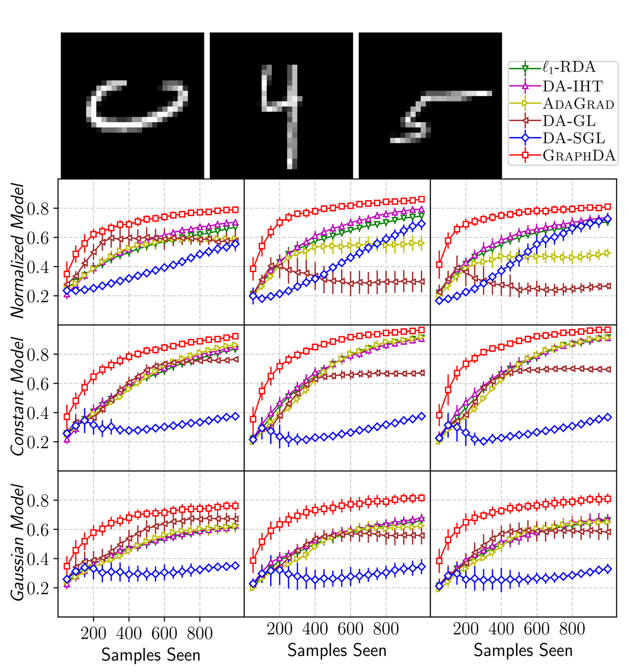


The learned models w_t of four benchmark graphs. For each pixel i, black stands for $(w_t)_i < 0$, gray $(w_t)_i = 0$, and white $(w_t)_i > 0$.



The classification accuracy as a function of number of training samples seen

- Online PGD-based: StoIHT and GraphStoIHT **do not work!**
- Online DA-based: ℓ_1-RDA,DA-GL,DA-SGL,DA-IHT work well.
- GraphDA outperforms other DA-based with the help of graph priors.



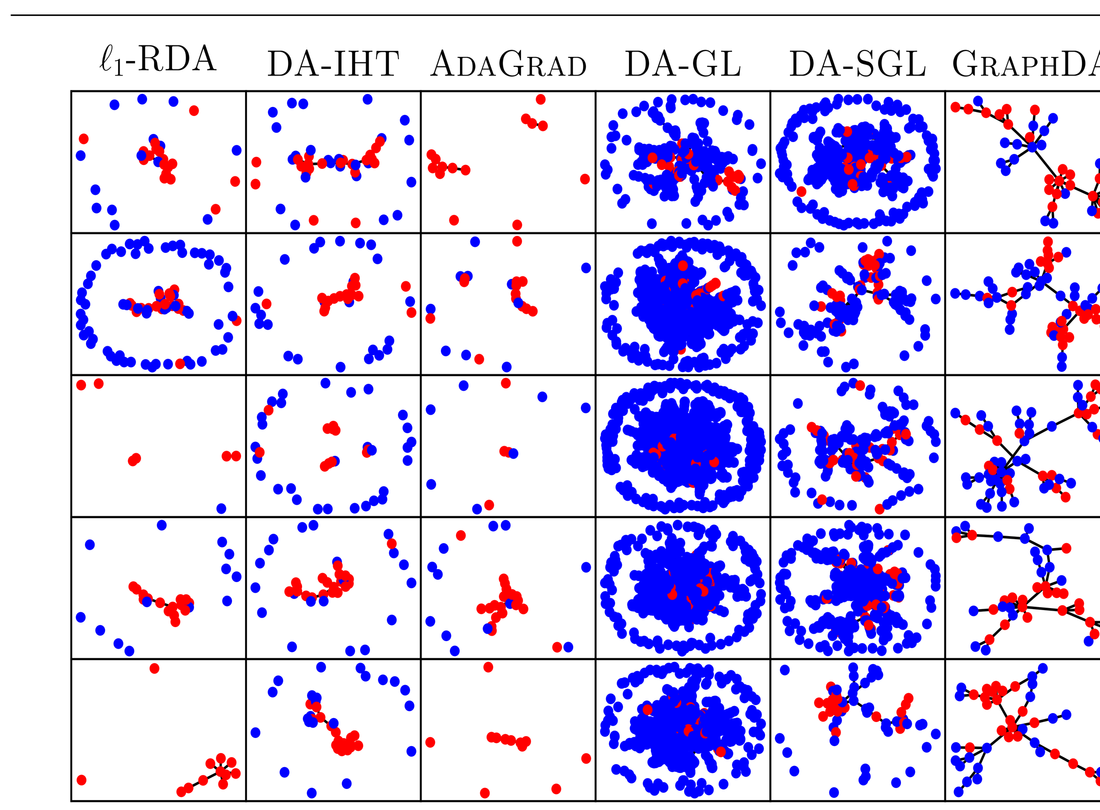
Online graph sparse linear regression: We generate 1,400 data samples by using the following linear relation: $y_t = \langle x_t, w^*\rangle$, where $x_t \in \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.

Three ways to generate w^*

- *Normalized Model*: each nonzero $w_i \in [0.0, 1.0]$
- *Constant Model*: each nonzero $w_i = 1.0$
- *Gaussian Model*: each nonzero $w_i \sim \mathcal{N}(0,1)$

Evaluation metric

$$F1_{w_t} = \frac{2|\mathrm{supp}(w^*) \cap \mathrm{supp}(w_t)|}{|\mathrm{supp}(w^*)| + |\mathrm{supp}(w_t)|}$$



• genes in HSA05213 • genes not in HSA05213

Conclusion and Future Work

Conclusion

- We propose a dual averaging-based method, GraphDA, for online graph-structured sparsity constraint problems.
- We prove that the minimization problem in the dual averaging step can be formulated as two equivalent optimization problems.
- GraphDA achieves better classification performance and stronger interpretability.

Future work

- Does GraphDA have non-regret bound under some proper assumption ?
- What if true structure of features are time evolving ?

Code & Datasets

- Code & Datasets can be found at GitHub: https://github.com/baojianzhou/graph-da
- Email: bzhou6@albany.edu
- **Baojian Zhou is open to postdoc positions.**