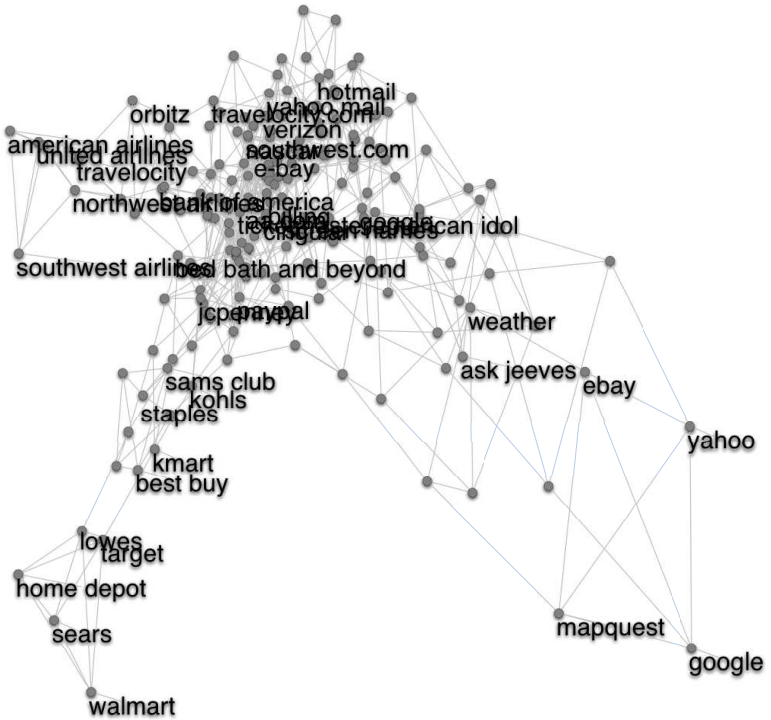
Laplacian Spectrum Learning

Pannaga Shivaswamy and **Tony Jebara**
Columbia University

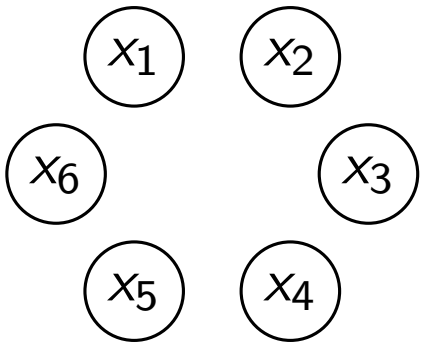
September 23, 2010

Learning with Graphs

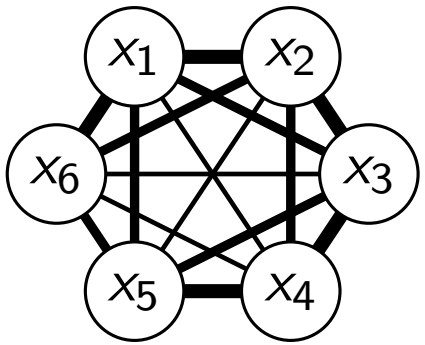


- Many learning problems are on graphs
- E.g. given labels of some nodes, predict others
- WWW, social networks, communication, trade networks...

Graph Laplacian in Machine Learning



Data



Weighted Graph

- Many machine learning methods convert data into a graph
- Given inputs $\mathbf{x}_1, \ldots, \mathbf{x}_l, \ldots, \mathbf{x}_n$ and labels $\mathbf{y} = (y_1 \ldots y_l)^\top$
- Compute a graph as weighted adjacency matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$, e.g. $\mathbf{W}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ or $\mathbf{W} =$ k-nearest-neighbors
- The Laplacian is $\mathbf{L} = \mathbf{D} - \mathbf{W}$ where $\mathbf{D}_{ij} = \delta_{ij} \sum_k \mathbf{W}_{ik}$

Graph Laplacian in Machine Learning

- Clustering
  - Spectral cut (Donath & Hoffman 73)
  - Normalized cut (Shi & Malik 00)
  - Spectral clustering (Ng, Jordan & Weiss 01)
  - Spectral graph partitioning (Joachims 03)
- Regularization and Diffusion
  - Kernel PCA (Schölkopf et al. 98)
  - Diffusion kernels (Kondor & Lafferty 02)
  - Laplacian regularization (Belkin & Niyogi 02)
  - Graph regularization (Smola & Kondor 03)
- Semi-supervised learning
  - Spectral graph transducer (Joachims 03)
  - Gaussian fields and harmonic functions (Zhu et al. 03)
  - Local and global consistency (Zhou et al. 04)
  - Laplacian support vector machines (Belkin et al. 06)
  - Graph transduction (Wang et al. 08)

Laplacian Regularization Properties

- The Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{W}$ regularizes functions on the graph. Function values on adjacent nodes can't be too different

$$\mathbf{f}^\top \mathbf{L}\mathbf{f} \;=\; \sum_{i<j} \mathbf{W}_{ij}(\mathbf{f}(i) - \mathbf{f}(j))^2$$

- Eigendecomposition is $\mathbf{L} = \sum_{i=1}^{n} \theta_i \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top$ where $\theta_i \leq \theta_{i+1}$
- Laplacian's bottom eigenvectors are smooth over the graph

$$\sum_{i<j} \mathbf{W}_{ij}(\boldsymbol{\phi}(i) - \boldsymbol{\phi}(j))^2 \;=\; \boldsymbol{\phi}^\top \mathbf{L}\boldsymbol{\phi} \;=\; \theta$$

- A *graph kernel* is built from the bottom q eigenvectors with *transformed* eigenvalues (Smola & Kondor 03)

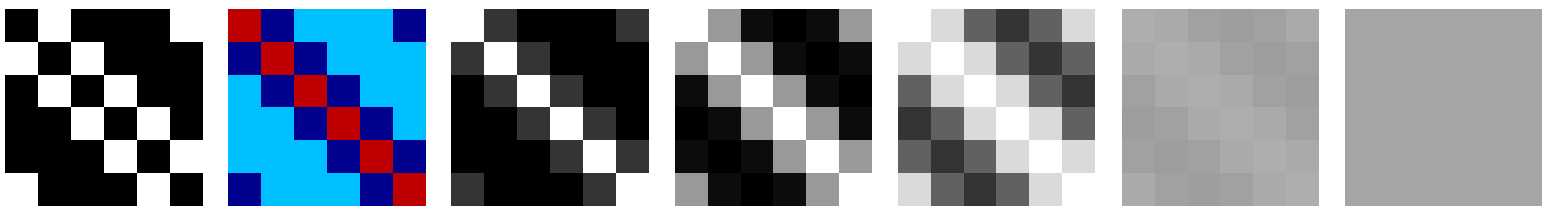
$$\mathbf{K} = \sum_{i=1}^{q} r(\theta_i)\boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top$$

Graph Kernels as Transformed Spectra: Diffusion

- To get a kernel from Laplacian, try various $r(\theta)$ functions

$$\mathbf{K} = \sum_{i=1}^{q} r(\theta_i)\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top$$

- Example: diffusion $r(\theta) = \exp(-\theta/\sigma^2)$



$\mathbf{W}$ $\mathbf{L}$ $\mathbf{K}_{\sigma=4}$ $\mathbf{K}_{\sigma=2}$ $\mathbf{K}_{\sigma=1}$ $\mathbf{K}_{\sigma=1/2}$ $\mathbf{K}_{\sigma=1/4}$

Graph Kernels as Transformed Spectra: Diffusion

- To get a kernel from Laplacian, try various $r(\theta)$ functions

$$\mathbf{K} = \sum_{i=1}^{q} r(\theta_i)\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top$$

- Example: diffusion $r(\theta) = \exp(-\theta/\sigma^2)$
- Example: kernel pca $r(\theta) = \frac{1}{\theta}\mathbf{I}(\theta < \sigma^2)$
- Example: Gaussian field kernel $r(\theta) = (\sigma^2 + \theta)^{-1}$
- *Any* monotonically decreasing $r(\theta)$ (Smola & Kondor 03)
- Searching over functions is tedious, can we learn $r(\theta)$?

Graph Kernels via Spectrum Learning

- Try max-margin multi-kernel learning (Lanckriet et al. 02)

$$\text{max margin s.t. } \mathcal{K} = \left\{ \mathbf{K} = \sum_{i=1}^{q} \mu_i \phi_i \phi_i^\top, \ \operatorname{tr}(\mathbf{K}) = 1, \ \mu_i \geq 0 \right\}$$

Problem: does not enforce spectral monotonicity

Graph Kernels via Spectrum Learning

- Try max-margin multi-kernel learning (Lanckriet et al. 02)

$$\text{max margin s.t. } \mathcal{K} = \left\{ \mathbf{K} = \sum_{i=1}^{q} \mu_i \phi_i \phi_i^\top, \ \operatorname{tr}(\mathbf{K}) = 1, \ \mu_i \geq 0 \right\}$$

  Problem: does not enforce spectral monotonicity

- Try kernel target alignment with monotonicity (Zhu et al. 04)

$$\max_{\boldsymbol{\mu}} \mathcal{A}(\mathbf{K}, \mathbf{y}\mathbf{y}^\top) \text{ s.t. } \mathbf{K} \in \mathcal{K}, \ \mu_i \geq \mu_{i+1}$$

  Problem: not max-margin, classifier is learned separately

Graph Kernels via Spectrum Learning

- Try max-margin multi-kernel learning (Lanckriet et al. 02)

$$\text{max margin s.t. } \mathcal{K} = \left\{ \mathbf{K} = \sum_{i=1}^{q} \mu_i \phi_i \phi_i^\top, \ \operatorname{tr}(\mathbf{K}) = 1, \ \mu_i \geq 0 \right\}$$

  Problem: does not enforce spectral monotonicity

- Try kernel target alignment with monotonicity (Zhu et al. 04)

$$\max_{\boldsymbol{\mu}} \mathcal{A}(\mathbf{K}, \mathbf{y}\mathbf{y}^\top) \text{ s.t. } \mathbf{K} \in \mathcal{K}, \ \mu_i \geq \mu_{i+1}$$

  Problem: not max-margin, classifier is learned separately

- Try max margin with spectral monotonicity (Xu et al. 07)

$$\text{max margin s.t. } \mathbf{K} \in \mathcal{K}, \ \mu_i \geq \mu_{i+1}$$

  Problem: flawed convex program! Let's fix this...

Absolute Margin and Relative Margin

- Instead of just max margin, consider a strict generalization...

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i \text{ s.t. } y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i$$

Absolute Margin and Relative Margin

- Instead of just max margin, consider a strict generalization...

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i \text{ s.t. } y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad |\mathbf{w}^\top \mathbf{x}_i + b| \leq B$$

- Relative margin machine (RMM) primal problem (S & J 08)

Absolute Margin and Relative Margin

- Instead of just max margin, consider a strict generalization...

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i \text{ s.t. } y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad |\mathbf{w}^\top \mathbf{x}_i + b| \leq B$$

- Relative margin machine (RMM) primal problem (S & J 08)

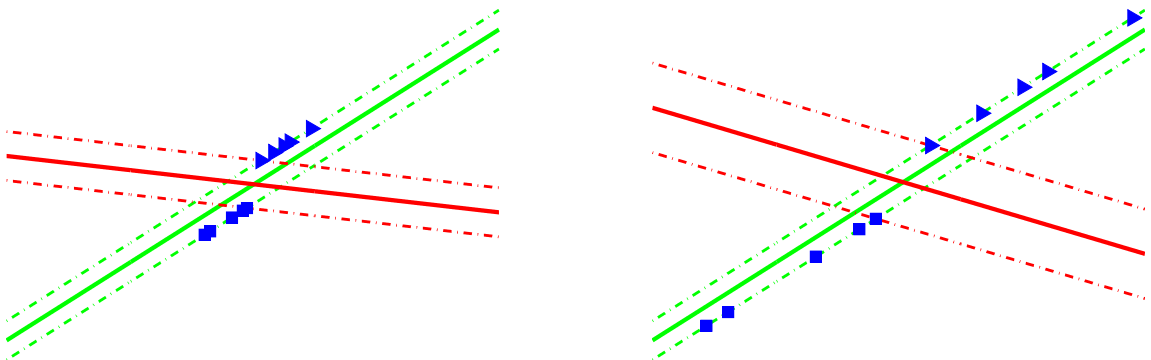


Figure: In red is the usual max-margin support vector machine ($B = \infty$) solution and in green is the max-relative margin ($B \approx 1$) solution.

Maximum Relative Margin with Spectrum Estimation

Start with relative margin machine dual problem (S & J 08)

$$\max_{(\boldsymbol{\alpha},\boldsymbol{\beta},\bar{\boldsymbol{\beta}})\in\Omega} \boldsymbol{\alpha}^\top \mathbf{1} - B(\boldsymbol{\beta}^\top \mathbf{1} + \bar{\boldsymbol{\beta}}^\top \mathbf{1}) - \frac{1}{2}\boldsymbol{\gamma}^\top \mathbf{K}\boldsymbol{\gamma}$$

where we have defined $\boldsymbol{\gamma} = \mathrm{diag}(\mathbf{y})\boldsymbol{\alpha} - \boldsymbol{\beta} + \bar{\boldsymbol{\beta}}$
and $\Omega : \left\{\boldsymbol{\alpha}^\top \mathbf{y} - \boldsymbol{\beta}^\top \mathbf{1} + \bar{\boldsymbol{\beta}}^\top \mathbf{1} = 0,\ \boldsymbol{\alpha}, \boldsymbol{\beta}, \bar{\boldsymbol{\beta}} \geq \mathbf{0},\ \boldsymbol{\alpha} \leq C\mathbf{1}\right\}$

Maximum Relative Margin with Spectrum Estimation

Start with relative margin machine dual problem (S & J 08)
Include monotonic optimization over the spectrum

$$\min_{\boldsymbol{\mu}} \max_{(\boldsymbol{\alpha},\boldsymbol{\beta},\bar{\boldsymbol{\beta}})\in\Omega} \boldsymbol{\alpha}^\top \mathbf{1} - B(\boldsymbol{\beta}^\top \mathbf{1} + \bar{\boldsymbol{\beta}}^\top \mathbf{1}) - \frac{1}{2}\boldsymbol{\gamma}^\top \mathbf{K}\boldsymbol{\gamma}$$

$$\text{s.t. } \mathbf{K} = \sum_{i=1}^{q} \mu_i \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top, \;\; \mathrm{tr}(\mathbf{K}) = 1, \;\; \mu_i \geq \mu_{i+1} \geq 0$$

Maximum Relative Margin with Spectrum Estimation

Start with relative margin machine dual problem (S & J 08)
Include monotonic optimization over the spectrum
Swap min and max (Boyd & Vandeberghe 04)

$$\max_{(\boldsymbol{\alpha},\boldsymbol{\beta},\bar{\boldsymbol{\beta}})\in\Omega} \min_{\boldsymbol{\mu}} \quad \boldsymbol{\alpha}^\top \mathbf{1} - B(\boldsymbol{\beta}^\top \mathbf{1} + \bar{\boldsymbol{\beta}}^\top \mathbf{1}) - \frac{1}{2}\boldsymbol{\gamma}^\top \mathbf{K}\boldsymbol{\gamma}$$

$$\text{s.t. } \mathbf{K} = \sum_{i=1}^{q} \mu_i \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top, \ \operatorname{tr}(\mathbf{K}) = 1, \ \mu_i \geq \mu_{i+1} \geq 0$$

Maximum Relative Margin with Spectrum Estimation

Consider inner minimization:

$$\min_{\boldsymbol{\mu}} \; -\frac{1}{2}\sum_{i=1}^{q} \boldsymbol{\gamma}^\top \mu_i \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top \boldsymbol{\gamma}$$

$$\text{s.t. } \mu_i - \mu_{i+1} \geq 0 \;,\; \mu_q \geq 0 \;,\; \mu_1 \geq 0 \;,\; \sum_{i=1}^{q} \mu_i = 1$$

Dual is:

$$\max_{\boldsymbol{\lambda}\geq \mathbf{0},\tau} \; -\tau$$

$$\text{s.t.} \quad \frac{1}{2}\boldsymbol{\gamma}^\top \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top \boldsymbol{\gamma} = -\lambda_i + \lambda_{i-1} + \tau$$

$$\text{s.t.} \quad \frac{1}{2}\boldsymbol{\gamma}^\top \boldsymbol{\phi}_1 \boldsymbol{\phi}_1^\top \boldsymbol{\gamma} = -\lambda_1 + \tau$$

Maximum Relative Margin with Spectrum Estimation

Consider inner minimization: (slight ϵ change)

$$\min_{\boldsymbol{\mu}} \; -\frac{1}{2}\sum_{i=1}^{q} \boldsymbol{\gamma}^\top \mu_i \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top \boldsymbol{\gamma}$$

$$\text{s.t. } \mu_i - \mu_{i+1} \geq \epsilon,\; \mu_q \geq \epsilon,\; \mu_1 \geq \epsilon,\; \sum_{i=1}^{q} \mu_i = 1$$

Dual is:

$$\max_{\boldsymbol{\lambda}\geq \mathbf{0},\tau} -\tau + \epsilon \sum_{i=1}^{q} \lambda_i$$

$$\text{s.t. } \sum_{j=2}^{i} \frac{1}{2} \boldsymbol{\gamma}^\top \boldsymbol{\phi}_j \boldsymbol{\phi}_j^\top \boldsymbol{\gamma} = -\lambda_i + \tau(i-1)$$

$$\text{s.t. } \frac{1}{2} \boldsymbol{\gamma}^\top \boldsymbol{\phi}_1 \boldsymbol{\phi}_1^\top \boldsymbol{\gamma} = -\lambda_1 + \tau$$

Maximum Relative Margin with Spectrum Estimation

Theorem: Given the following max-min problem,

$$\max_{(\boldsymbol{\alpha},\boldsymbol{\beta},\bar{\boldsymbol{\beta}})\in\Omega} \min_{\boldsymbol{\mu}} \boldsymbol{\alpha}^\top \mathbf{1} - B(\boldsymbol{\beta}^\top \mathbf{1} + \bar{\boldsymbol{\beta}}^\top \mathbf{1}) - \frac{1}{2}\boldsymbol{\gamma}^\top \mathbf{K}\boldsymbol{\gamma}$$

$$\text{s.t. } \mathbf{K} = \sum_{i=1}^{q} \mu_i \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top, \ \operatorname{tr}(\mathbf{K}) = 1, \ \mu_i \geq \mu_{i+1} \geq 0$$

Maximum Relative Margin with Spectrum Estimation

Theorem: Given the following max-min problem,

$$\max_{(\boldsymbol{\alpha},\boldsymbol{\beta},\bar{\boldsymbol{\beta}})\in\Omega} \min_{\boldsymbol{\mu}} \boldsymbol{\alpha}^\top \mathbf{1} - B(\boldsymbol{\beta}^\top \mathbf{1} + \bar{\boldsymbol{\beta}}^\top \mathbf{1}) - \frac{1}{2}\boldsymbol{\gamma}^\top \mathbf{K}\boldsymbol{\gamma}$$

$$\text{s.t. } \mathbf{K} = \sum_{i=1}^{q} \mu_i \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top, \ \operatorname{tr}(\mathbf{K}) = 1, \ \mu_i \geq \mu_{i+1} \geq 0$$

a nearly equivalent optimization (for small ϵ) is

$$\max_{(\boldsymbol{\alpha},\boldsymbol{\beta},\bar{\boldsymbol{\beta}})\in\Omega} \max_{\boldsymbol{\lambda}\geq\mathbf{0},\tau} \boldsymbol{\alpha}^\top \mathbf{1} - B(\boldsymbol{\beta}^\top \mathbf{1} + \bar{\boldsymbol{\beta}}^\top \mathbf{1}) - \tau + \epsilon \sum_{i=1}^{q} \lambda_i$$

$$\text{s.t. } \sum_{j=2}^{i} \frac{1}{2}\boldsymbol{\gamma}^\top \boldsymbol{\phi}_j \boldsymbol{\phi}_j^\top \boldsymbol{\gamma} \leq -\lambda_i + \tau(i-1) \quad i = 2, \ldots, q$$

$$\text{s.t. } \frac{1}{2}\boldsymbol{\gamma}^\top \boldsymbol{\phi}_1 \boldsymbol{\phi}_1^\top \boldsymbol{\gamma} \leq -\lambda_1 + \tau.$$

Maximum Relative Margin with Monotonic Transformation

- End up with this final QCQP which requires $O(ql^3)$

$$\max_{(\boldsymbol{\alpha},\boldsymbol{\beta},\bar{\boldsymbol{\beta}})\in\Omega} \max_{\boldsymbol{\lambda}\geq\mathbf{0},\tau} \boldsymbol{\alpha}^\top\mathbf{1} - B(\boldsymbol{\beta}^\top\mathbf{1} + \bar{\boldsymbol{\beta}}^\top\mathbf{1}) - \tau + \epsilon\sum_{i=1}^{q}\lambda_i$$

$$\text{s.t. } \sum_{j=2}^{i}\frac{1}{2}\boldsymbol{\gamma}^\top\boldsymbol{\phi}_j\boldsymbol{\phi}_j^\top\boldsymbol{\gamma} \leq -\lambda_i + \tau(i-1) \quad i = 2,\ldots,q$$

$$\text{s.t. } \frac{1}{2}\boldsymbol{\gamma}^\top\boldsymbol{\phi}_1\boldsymbol{\phi}_1^\top\boldsymbol{\gamma} \leq -\lambda_1 + \tau$$

where $\boldsymbol{\gamma} = \mathrm{diag}(\mathbf{y})\boldsymbol{\alpha} - \boldsymbol{\beta} + \bar{\boldsymbol{\beta}}$ and
$\Omega : \left\{\boldsymbol{\alpha}^\top\mathbf{y} - \boldsymbol{\beta}^\top\mathbf{1} + \bar{\boldsymbol{\beta}}^\top\mathbf{1} = 0,\ \boldsymbol{\alpha},\boldsymbol{\beta},\bar{\boldsymbol{\beta}} \geq \mathbf{0},\ \boldsymbol{\alpha} \leq C\mathbf{1}\right\}.$

- Setting $B = \infty$ gives the maximum absolute margin solution
- STORM: Spectral Transformations that Optimize the Relative Margin
- STOAM: Spectral Transformations that Optimize the Absolute Margin

Experiments - Methods

- Use a semi-supervised learning experimental setting
- The number of labeled examples is varied from 30 to 110
- The remaining examples used as unlabeled data
- The following algorithms were compared:

Xu07	Flawed version of STOAM (Xu et al. 07)
MKL-S	(Lanckriet et al. 02) with max margin
MKL-R	(Lanckriet et al. 02) with max relative margin
SGT	Spectral graph transducer (Joachims 03)
KTA-S	(Zhu et al. 04) followed by an SVM
KTA-R	(Zhu et al. 04) followed by an RMM
STOAM	Spectral transformations that optimize absolute margin
STORM	Spectral transformations that optimize relative margin

Experiments - Mean error rates on text datasets

	l	Xu07	MKL-S	MKL-R	SGT	KTA-S	KTA-R	STOAM	STORM
r-a	30	44.89	37.14	37.14	19.46	22.98	22.99	25.81	25.81
	110	37.74	20.43	20.41	18.10	16.46	16.46	16.40	16.41
w-m	30	46.98	22.74	22.74	41.88	16.03	16.08	14.26	14.26
	110	41.91	11.84	11.85	18.16	10.87	10.99	10.31	10.28
p-m	30	46.48	41.21	40.99	39.58	28.00	28.05	30.58	30.58
	110	38.10	25.88	26.16	32.16	19.53	19.56	19.74	19.70
b-h	30	47.04	4.35	4.35	3.95	3.91	3.80	3.90	3.87
	110	44.99	3.88	3.88	3.83	3.71	3.66	3.67	3.56
m-m	30	48.11	12.35	12.35	41.30	7.35	7.36	7.60	6.88
	110	41.94	5.44	5.16	10.96	4.97	4.92	4.95	4.65

Table: In each row, minimum error rate is in red. Algorithms whose performance is not significantly worse (at 5% significance) are in blue.

Experiments - Mean error rates on digit datasets

	l	Xu07	MKL-S	MKL-R	SGT	KTA-S	KTA-R	STOAM	STORM
0-9	30	46.45	0.89	0.89	0.83	0.90	0.90	0.88	0.88
	110	45.40	0.85	0.90	0.87	0.89	0.89	0.92	0.86
1-2	30	47.22	3.39	4.06	11.81	2.92	2.92	2.88	2.85
	110	44.97	2.77	2.36	2.61	2.61	2.61	2.51	2.51
3-8	30	45.42	13.02	12.63	9.86	8.54	7.58	7.93	7.68
	110	41.09	6.56	6.15	6.91	5.35	5.43	5.25	5.24
4-7	30	44.85	5.74	5.54	5.60	4.27	4.09	3.64	3.57
	110	41.85	3.28	3.00	3.60	2.99	2.98	2.92	2.91
5-6	30	46.75	5.18	4.91	2.49	3.48	3.32	3.19	2.96
	110	43.59	2.62	2.52	2.51	2.57	2.53	2.55	2.49

Table: In each row, minimum error rate is in red. Algorithms whose performance is not significantly worse (at 5% significance) are in blue.

Experiments - Summary

- Adjacency matrix **W** was a 5-nearest-neighbor graph
- Hyperparameters set to $\epsilon = 10^{-6}$ and $q = 200$ throughout
- Parameters C and B found via cross-validation
- Summary of MNIST digits and Newsgroups text results:

	Xu07	MKL-S	MKL-R	SGT	KTA-S	KTA-R	STOAM	STORM
#best	0	1	5	9	5	2	8	22
#O(best)	0	1	2	4	8	12	16	13
#total	0	2	7	13	13	14	24	35

Table: For each method, we enumerate the number of times it performed best, the number of times it was not significantly worse than best and the total number of times it was either best or not significantly worse.

Experiments - Spectrum Visualization on Text Data

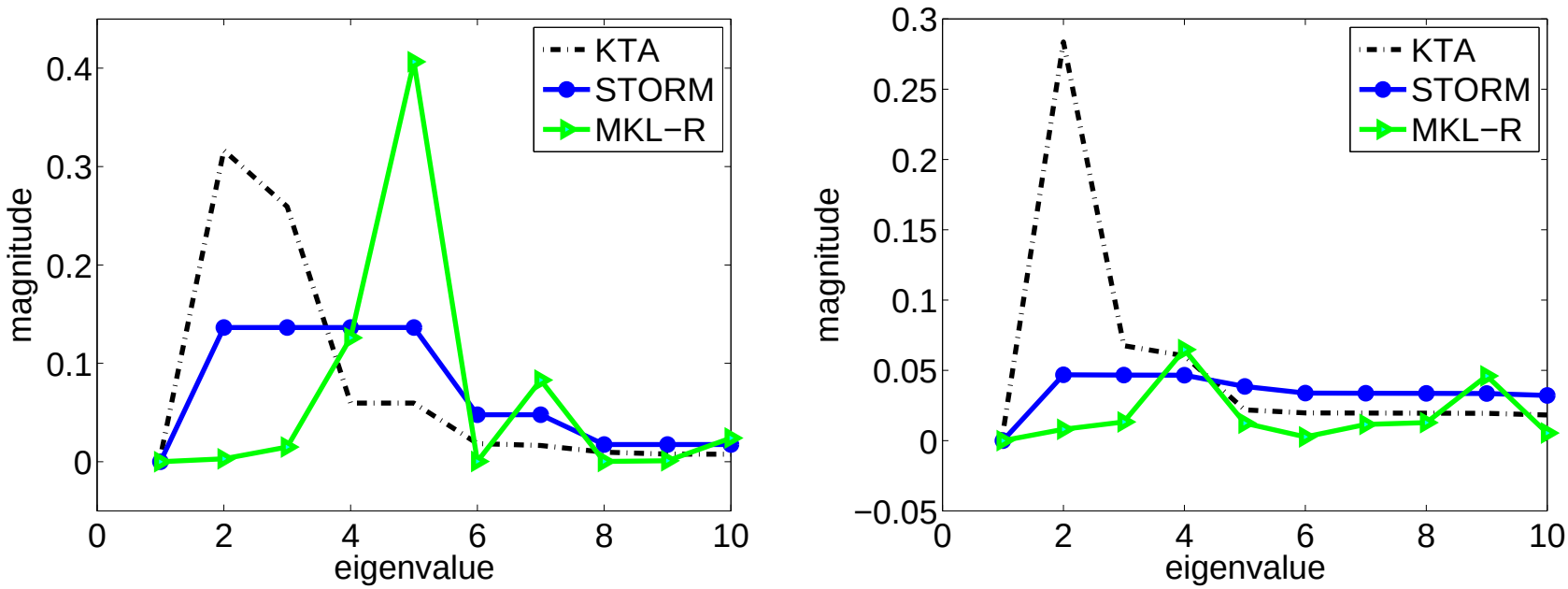


Figure: Magnitudes of the top eigenvalues recovered by the different algorithms for problems m-m and p-m. The plots show average eigenspectra over all experiments. KTA and STORM have monotonically decreasing spectra.

Experiments - Spectrum Visualization on Digits Data

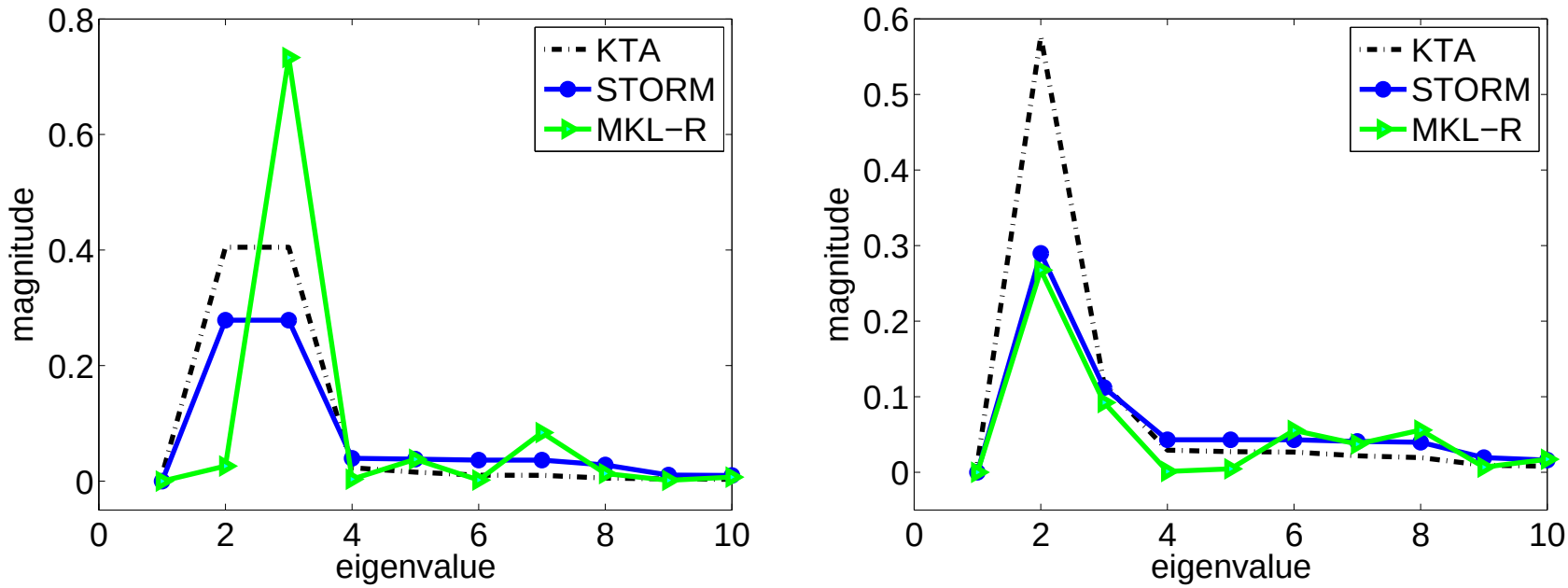


Figure: Magnitudes of the top eigenvalues recovered by the different algorithms for problems 1-2 and 3-8. The plots show average eigenspectra over all experiments. KTA and STORM have monotonically decreasing spectra.

Conclusions

- Implemented general graph kernel semi-supervised learning
- Explore any monotonic transformation of Laplacian spectrum
- Simultaneously learn kernel and classifier
- Both optimized under large margin criterion
- Better yet, both optimized under relative margin criterion
- All computations are efficient convex programs
- Significantly outperforms other kernel-learning methods