Data-Dependent Geometries and Structures : Analyses and Algorithms for Machine Learning

Mark Herbster, Guy Lever, John Shawe-Taylor
University College London
m.herbster@cs.ucl.ac.uk g.lever@cs.ucl.ac.uk
jst@cs.ucl.ac.uk

Claudio Gentile, Fabio Vitale
Universita' dell'Insubria, Varese
claudio.gentile@uninsubria.it,
fabiovdk@yahoo.com

Nello Cristianini
University of Bristol
nello.cristianini@gmail.com

29th March 2012

Data Dependent Geometry

What is a *"data-dependent geometry"*?

Standard paradigm

- A dataset is sampled from a space with a ***given*** geometry
- the "distances" between particular points is ***independent*** of the sample

Data-dependent paradigm

- A dataset is sampled from a space with an ***unknown*** geometry
- Hence the "distances" between particular points is ***dependent*** on the sample
- **Implication:** We need to learn the "geometry" (Assumptions Needed!)

Data Dependent Geometry

What is a *"data-dependent geometry"*?

Standard paradigm

- A dataset is sampled from a space with a ***given*** geometry
- the "distances" between particular points is ***independent*** of the sample

Data-dependent paradigm

- A dataset is sampled from a space with an ***unknown*** geometry
- Hence the "distances" between particular points is ***dependent*** on the sample
- **Implication:** We need to learn the "geometry" (Assumptions Needed!)

Example

Consider the following dataset of a new stories

News stories (Source, Headline)

1. (*Financial Times*, Research and Development in Fusion increased by 60% Last Quarter)
2. (*St. Petersburg Gazeteer*, Major layoffs expected in tourism sector)
3. (*The Times*, Super-Tanker founders on Florida coast. Largest spill of the millennium.)

Observation

Knowing "3" suggests the distance from "1" and "2" be reduced

Example

Consider the following dataset of a new stories

News stories (Source, Headline)

1. (*Financial Times*, Research and Development in Fusion increased by 60% Last Quarter)
2. (*St. Petersburg Gazeteer*, Major layoffs expected in tourism sector)
3. (*The Times*, Super-Tanker founders on Florida coast. Largest spill of the millennium.)

Observation

Knowing “3” suggests the distance from “1” and “2” be reduced

Example

Consider the following dataset of a new stories

News stories (Source, Headline)

1. (*Financial Times*, Research and Development in Fusion increased by 60% Last Quarter)
2. (*St. Petersburg Gazeteer*, Major layoffs expected in tourism sector)
3. (*The Times*, Super-Tanker founders on Florida coast. Largest spill of the millennium.)

Observation

Knowing “3” suggests the distance from “1” and “2” be reduced

Example

Consider the following dataset of a new stories

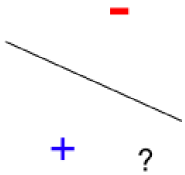
News stories (Source, Headline)

1. (*Financial Times*, Research and Development in Fusion increased by 60% Last Quarter)
2. (*St. Petersburg Gazeteer*, Major layoffs expected in tourism sector)
3. (*The Times*, Super-Tanker founders on Florida coast. Largest spill of the millennium.)

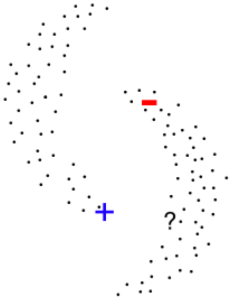
Observation

Knowing “3” suggests the distance from “1” and “2” be reduced

Illustration



Illustration



Data-dependent Geometry

Topics

- Graph-based semi-supervised learning
  - Laplacian-based methods (Data dependent kernels)
  - Tree approximations (online mistake bounds)
  - Link classification (Active learning)
  - Fast algorithms (Bayesian Marginalisation)
- Exploiting the structure of an unknown data-generating distribution
  - Localized Pac-Bayes analysis

Resources Allocated

Resources

Activity	**duration**	**cost**
Guy Lever **RA** (UCL)	5 months	€23K
Fabio Vitale **RA** (Insubria)	9 months	€19K
Travel and subsistence	—	€3K
Total:	—	€45K

Outputs

1. N. Cesa-Bianchi, C. Gentile, F. Vitale, and G. Zappella. A correlation clustering approach to link classification in signed networks., *Submitted*, 2012.
2. N. Cesa-Bianchi, C. Gentile, F. Vitale, and G. Zappella. See the tree through the lines: the shazoo algorithm., *NIPS*, 2012.
3. M. Herbster. A triangle inequality for p-resistance., *NIPS Workshop: Networks Across Disciplines: Theory and Applications*, 2010.
4. M. Herbster, S. Pasteris, and F. Vitale. Efficient prediction for tree markov random fields in a streaming model., *NIPS Workshop on Discrete Optimization in Machine Learning*, 2011.
5. G. Lever, T. Diethe, and J. Shawe-Taylor. Data dependent kernels in nearly-linear time., *AISTATS*, 2012.
6. G. Lever, F. Laviolette, and J. Shawe-Taylor. Tighter pac-bayes bounds through distribution-dependent priors., *Theoretical Computer Science (To appear)*, 2012.

Main Insubria activities

- **Vertex classification on weighted graphs**

N. Cesa-Bianchi, C. Gentile, F. Vitale, and G. Zappella. See the tree through the lines: the shazoo algorithm. In Proc. of 25th NIPS, 2012.

- **Link classification on unweighted graphs**

N. Cesa-Bianchi, C. Gentile, F. Vitale, and G. Zappella. A correlation clustering approach to link classification in signed networks. Submitted, 2012.

- **Main issues**:
  - Construction of meaningful and natural **complexity measures**
  - Accuracy guarantees / **optimality**
  - **Scalability**
  - **Practical utility**

- **Performance measure (analysis):** number of prediction mistakes

Vertex Classification
The Shazoo algorithm

- **Learning on graphs/trees domains**: hyperlinked webpages, social networks, co-author networks, biological networks, …
- **Our learning problem: Vertex classification** of weighted, connected and undirected **trees (and graphs)** based only on **graph topology**
- We focus on **binary labeling**
- **Bias: strongly connected nodes → same label**
  Weight cut-edges **small**

The Shazoo algorithm *[Cesa-Bianchi et al. NIPS 2012]:* **input = weighted trees T**
(if the input is a graph G we can run Shazoo on a **spanning tree** T of G)

- **Shazoo (1) partitions** T into components (satifying some properties), **(2)** uses **mincut** for estimating the labels of the component **border** vertices, **(3)** uses a **NN method** for predicting the required label

N. Cesa-Bianchi, C. Gentile, F. Vitale, and G. Zappella. See the tree through the lines: the shazoo algorithm. In Proc. of 25th NIPS, 2012.

Shazoo Algortihm: Analysis, implementation and computational complexity

Accuracy: **#mistakes** of Shazoo is **optimal** (up to log factors)

Implementation: simple and fast recursive method (based on sum-product algorithm) for using the mincut strategy

Time complexity:

- **On line protocol:** Worst case time per prediction: ***O*** **(#vertices)** (rarely encountered in practice)

- **Batch protocol** (vertices are split into training and test sets) : Worst case time for predicting **all** labels of the test set: ***O*** **(#vertices)**

Space complexity: Linear in #vertices

N. Cesa-Bianchi, C. Gentile, F. Vitale, and G. Zappella. See the tree through the lines: the shazoo algorithm. In Proc. of 25th NIPS, 2012.

Shazoo algorithm
Experiments

- **Real-world weighted graphs**: web spam detection, character recognition, text categorization and bioinformatics
- **Competitors: LABPROP** (label propagation algorithm), **OMV** (label majority vote of adjacent nodes) and **WTA** (Weighted Tree Algorithm)
- **We used spanning trees** generated in different ways **for running Shazoo (and WTA)**
- **Experiment protocol**: **batch** (training set size = **5%, 10%** and **25%)**
- **Main results:**
  - **Shazoo outperforms WTA and OMV on all datasets**

    (unlike **WTA** it **explicity exploits the tree structure**)
  - Aggregating prediction of committees of random spanning trees via majority vote, **Shazoo outperforms LABPROP** when the training set size is small

N. Cesa-Bianchi, C. Gentile, F. Vitale, and G. Zappella. See the tree through the lines: the shazoo algorithm. In Proc. of 25th NIPS, 2012.

Link classification

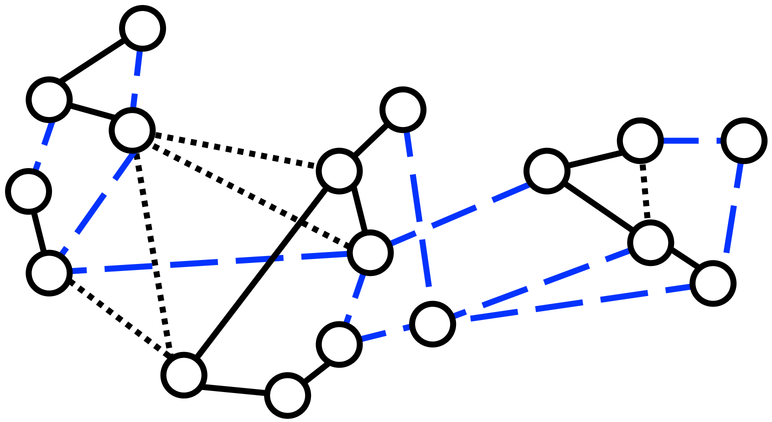
Protocol: Active Learning (focus)

Negative edges in real world networks:

Disapproval or distrust in social networks, negative endorsements on the Web, inhibitory interactions in biological networks, sentiment between two individuals for recommender systems

Active learning protocol

- Learner selects a set TrSet of edges (training set)
- All labels of the edges of TrSet are revealed
- Learner predicts the labels of all remaining edges



Edge label +1 → similarity

Edge label -1 → dissimilarity

Hidden label

N. Cesa-Bianchi, C. Gentile, F. Vitale, and G. Zappella. A correlation clustering approach to link classification in signed networks. Submitted, 2012.

Active link classification
Main results

- For this problem we studied a meaningful and natural **complexity measure** related to a notion of cutsize induced by Correlation Clustering
- **Accuracy guarantees:** We devised an algorithm **optimal** up to a $O\left(\rho^{3/2}\sqrt{|V|}\right)$ factor on any labeled graph **G(V,E)**, while the test set size is not smaller than ρ times the training test size
- **Scalability:** Our algorithm requires an amortized time per prediction equal to $O\left(\sqrt{\frac{|V|}{\rho}}\log|V|\right)$
- Research directions:
  - Use **randomization** against adversarial label assignment
  - Test our algorithm on real-world graphs drawn from different domains: social networks (Epinions, Slashdot), movie rating datasets (Movielens) and other web datasets (political election datasets, …)

N. Cesa-Bianchi, C. Gentile, F. Vitale, and G. Zappella. A correlation clustering approach to link classification in signed networks. Submitted, 2012.

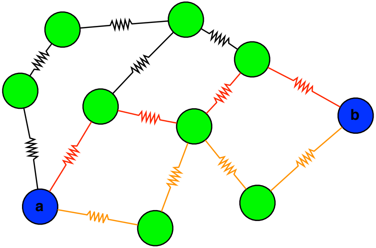
UCL Activities (Part I)

- Exploiting the structure of a graph (resistance metric)
- Fast online algorithms for labeling a graph

1. A triangle inequality for p-resistance.
    - p-resistance generalises the effective resistance of a network
    - Laplacian and Mincut methods popular, p-resistance for SSL generalises both
    - Fundamental inequality for p-resistance
    - Geometric insight given for k-center clustering
2. Efficient prediction for tree markov random fields in a streaming model
    - Exponential speedup for online tree MRF vertex marginalization
    - Computational complexity – characterised by a particular hierarchal covering of a tree

A triangle inequality for p-resistance (1)

1. Identify a graph with a network of resistors



2. **Definition:** The (effective) p-resistance from a and b is

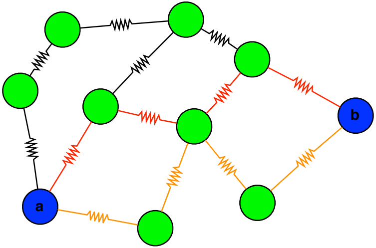
$$r_p(a,b) = \left[\min_{\boldsymbol{u} \in \mathbb{R}^n} \left\{ \sum_{(i,j) \in E(\mathbf{G})} \frac{|u_i - u_j|^p}{\pi_{ij}} : u_a = 1, u_b = 0 \right\} \right]^{-1}$$

3. p-Resistance trades off geodesic distance and connectivity
4.

Resistors in parallel	Resistors in series
$r_p^{\text{par}}(a,b) = \left(\sum_{i=1}^{n} \frac{1}{\pi_i} \right)^{-1}$	$r_p^{\text{ser}}(a,b) = \left(\sum_{i=1}^{n} \pi_i^{\frac{1}{p-1}} \right)^{p-1}$

A triangle inequality for p-resistance (1)

1. Identify a graph with a network of resistors



2. **Definition:** The (effective) p-resistance from a and b is

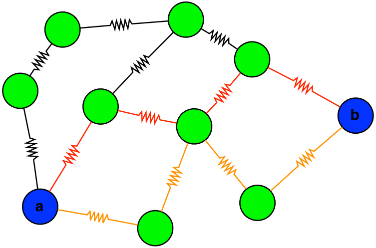
$$r_p(a,b) = \left[\min_{\boldsymbol{u} \in \mathbb{R}^n} \left\{ \sum_{(i,j) \in E(\mathbf{G})} \frac{|u_i - u_j|^p}{\pi_{ij}} : u_a = 1, u_b = 0 \right\} \right]^{-1}$$

3. p-Resistance trades off geodesic distance and connectivity
4. Resistors in parallel — Resistors in series

$$r_p^{\text{par}}(a,b) = \left(\sum_{i=1}^{n} \frac{1}{\pi_i} \right)^{-1} \qquad r_p^{\text{ser}}(a,b) = \left(\sum_{i=1}^{n} \pi_i^{\frac{1}{p-1}} \right)^{p-1}$$

A triangle inequality for p-resistance (1)

1. Identify a graph with a network of resistors



2. **Definition:** The (effective) p-resistance from a and b is

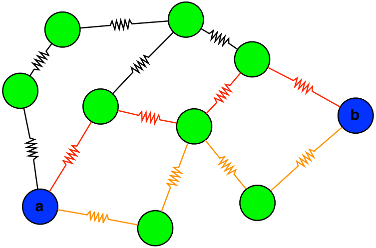
$$r_p(a,b) = \left[\min_{\boldsymbol{u} \in \mathbb{R}^n} \left\{ \sum_{(i,j) \in E(\mathbf{G})} \frac{|u_i - u_j|^p}{\pi_{ij}} : u_a = 1, u_b = 0 \right\} \right]^{-1}$$

3. p-Resistance trades off geodesic distance and connectivity
4. Resistors in parallel | Resistors in series

$$r_p^{\text{par}}(a,b) = \left(\sum_{i=1}^{n} \frac{1}{\pi_i} \right)^{-1} \qquad r_p^{\text{ser}}(a,b) = \left(\sum_{i=1}^{n} \pi_i^{\frac{1}{p-1}} \right)^{p-1}$$

A triangle inequality for p-resistance (1)

1. Identify a graph with a network of resistors



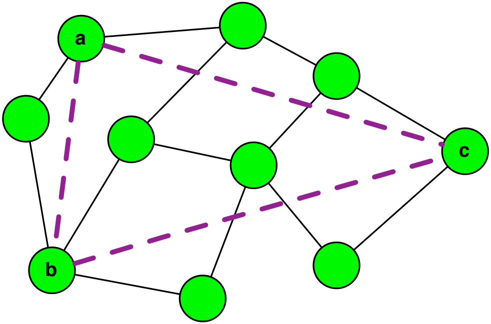
2. **Definition:** The (effective) p-resistance from a and b is

$$r_p(a,b) = \left[\min_{\boldsymbol{u} \in \mathbb{R}^n} \left\{ \sum_{(i,j) \in E(\mathbf{G})} \frac{|u_i - u_j|^p}{\pi_{ij}} : u_a = 1, u_b = 0 \right\} \right]^{-1}$$

3. p-Resistance trades off geodesic distance and connectivity
4.

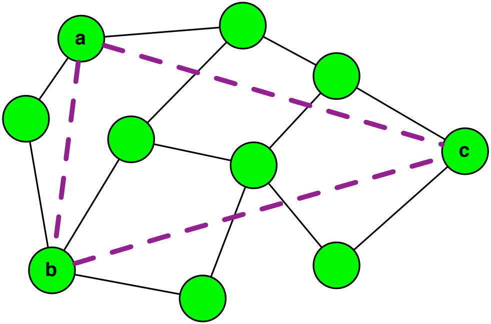
Resistors in parallel	Resistors in series
$r_p^{\text{par}}(a,b) = \left(\sum_{i=1}^{n} \frac{1}{\pi_i} \right)^{-1}$	$r_p^{\text{ser}}(a,b) = \left(\sum_{i=1}^{n} \pi_i^{\frac{1}{p-1}} \right)^{p-1}$

A triangle inequality for p-resistance (2)



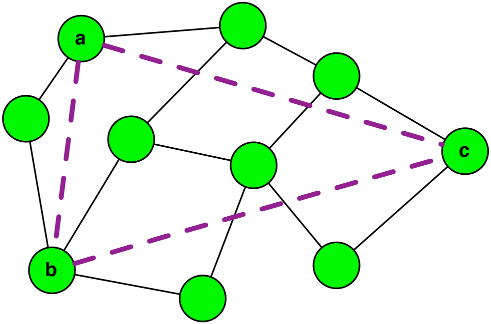
1. Electric Network ($p = 2$): $r_2(a, c) \le r_2(a, b) + r_2(b, c)$
2. Pipe Network ($p = 1$): $r_1(a, c) \le \max(r_1(a, b), r_1(b, c))$
3. Generic $p \in (1, \infty)$: $r_p(a, c) \le \left(r_p(a, b)^{\frac{1}{p-1}} + r_p(b, c)^{\frac{1}{p-1}} \right)^{p-1}$

A triangle inequality for p-resistance (2)



1. Electric Network ($p = 2$): $r_2(a,c) \leq r_2(a,b) + r_2(b,c)$
2. Pipe Network ($p = 1$): $r_1(a,c) \leq \max(r_1(a,b), r_1(b,c))$
3. Generic $p \in (1, \infty)$: $r_p(a,c) \leq \left(r_p(a,b)^{\frac{1}{p-1}} + r_p(b,c)^{\frac{1}{p-1}} \right)^{p-1}$

A triangle inequality for p-resistance (2)



1. Electric Network ($p = 2$): $r_2(a,c) \leq r_2(a,b) + r_2(b,c)$
2. Pipe Network ($p = 1$): $r_1(a,c) \leq \max(r_1(a,b), r_1(b,c))$
3. Generic $p \in (1,\infty)$: $r_p(a,c) \leq \left(r_p(a,b)^{\frac{1}{p-1}} + r_p(b,c)^{\frac{1}{p-1}} \right)^{p-1}$

A triangle inequality for p-resistance (3)

Application: k-center clustering

Objective:

$$\min_{v_1^*,\ldots,v_k^* \in V} \max_{v \in V} \min_{i \in \mathbb{N}_k} d(v, v_i^*) \,.$$

Farthest first algorithm

Input: A set $V = v_1, \ldots, v_n$, a $k \in \mathbb{N}$, and a metric $d(V, V) \to \mathbb{R}$
Initialization: $\tilde{v}_1 = v_1$
for $t = 2, \ldots, k$ **do**
$\tilde{v}_t = \text{argmax}_{v \in V} \min_{i \in \mathbb{N}_{t-1}} d(v, \tilde{v}_i)$
end for
return $\{\tilde{v}_1, \ldots, \tilde{v}_k\}$

Theorem

Given a graph $\mathcal{G}$ the farthest first algorithm gives a 2^{p-1}-opt k-center clustering with respect to the p-resistance for $p > 1$.

Efficient prediction for tree markov random fields (1)

Model

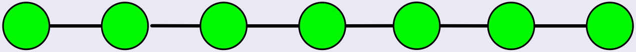
Given a tree-structured MRF at time $t = 1, 2, \ldots$

Actions:

i) predict a label at a vertex on the tree
ii) update by associating a label with a vertex
iii) delete the label at a vertex.

Problem

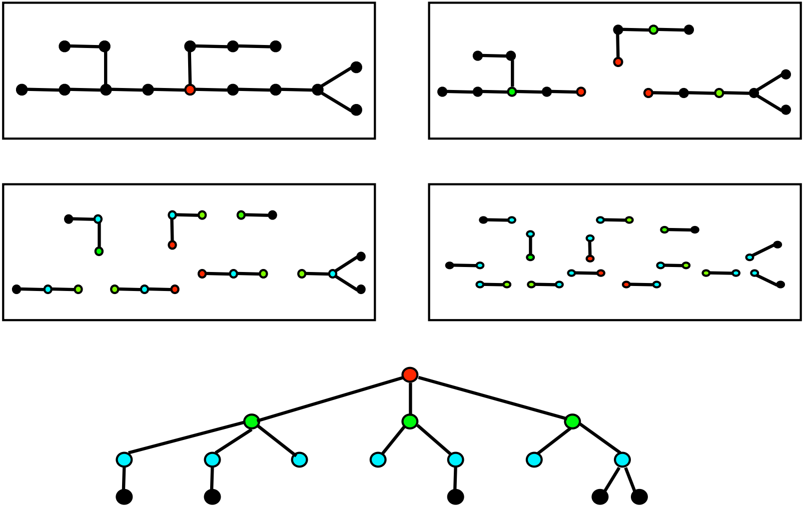
Problem: Online belief propagation is *slow* — linear on a tree.



Solution: We construct a (*decomposition*) tree on the original
Result: D-propagation is **fast** on a tree.

Efficient prediction for tree markov random fields (2)

Decompose the tree...



D-propagation

We construct tree D from T of height χ s.t.

$$\log(\text{height}(T)) \leq \chi \leq \min(\log(|T|), \text{height}(T)).$$

For update and prediction we then "D-propagate" on D.

	Online belief propagation	Online D-propagation
Prediction	$O(1)$	$O(\chi)$
Update	$O(\|T\|)$	$O(\chi)$
Initialisation	$O(\|T\|)$	$O(\|T\|^3)$ now $O(T)$

UCL Activities (Part II)

- Learning with data-dependent hypothesis classes
- Theoretical and practical advances
- 2 papers:
  1. Data dependent kernels in nearly-linear time
     - kernels on general (continuous) spaces capture data-defined structure
     - current methods scale poorly
     - exploit huge amounts of data
     - practical, fast
  2. Tighter PAC-Bayes bounds through distribution dependent priors
     - bounds for exponential weights and SVMs
     - Localized PAC-Bayes analysis
     - encode assumptions about interaction of classifiers with data
     - tight bounds, new distribution-dependent complexity measure

Data dependent kernels in nearly-linear time (1)

- kernels on general (continuous) domains capture structure in data
  - manifold structure, cluster structure etc.
- we want:
  - Fast (need to exploit lots of data to be robust)
  - automatic (no tuning or domain knowledge)
- Problem: Given space $\mathcal{X}$ and subsample $\mathcal{V} \subset \mathcal{X}$, $|\mathcal{V}| = n$ and "intrinsic regularizer":

$$\mathrm{reg}(h) = \boldsymbol{h}^\top \boldsymbol{Q} \boldsymbol{h} \tag{1}$$

where $h : \mathcal{X} \to \mathbb{R}$ and $h_i = h(v_i)$, define kernel $\widetilde{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that:
  - functions $h \in \mathcal{H}_{\widetilde{K}}$ smooth w.r.t. (1)
  - $\widetilde{K}$ extends kernel $\boldsymbol{Q}^+$ from $\mathcal{V}$ to $\mathcal{X}$

Data dependent kernels in nearly-linear time (2)

- One solution (Sindhwani et. al. 2005): pick basic $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ then define

$$\langle h, g \rangle_{\widetilde{K}} := \beta \langle h, g \rangle_K + (1-\beta) \boldsymbol{h}^\top \boldsymbol{Q} \boldsymbol{g}, \quad h, g \in \mathcal{H}_K, \tag{2}$$

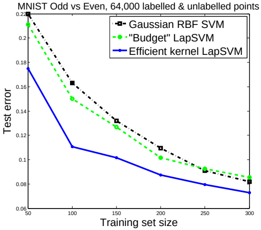
- kernel $\widetilde{K}$ has closed form, but cubic complexity
- solution: disconnect $\mathcal{V}$ from *landmark points* $\mathcal{L} \subset \mathcal{V}$ at which functions in $\mathcal{H}_K$ are measured
- Proposed RKHS has inner probuct:

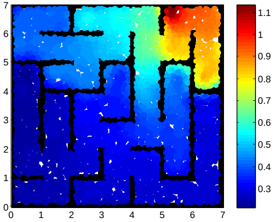
$$\langle h, g \rangle_{\breve{K}} := \beta \langle h, g \rangle_K + (1-\beta) (\boldsymbol{h}^*)^\top \boldsymbol{Q} \boldsymbol{g}^*, \quad h, g \in \mathcal{H}_K, \tag{3}$$

  where $\boldsymbol{h}|_{\mathcal{L}}$ is restriction of $\boldsymbol{h} \in \mathbb{R}^{\mathcal{V}}$ to $\mathcal{L}$,
  $\boldsymbol{h}^* \in \text{argmin}_{\boldsymbol{h} \in \mathbb{R}^{\mathcal{V}}} \{ \boldsymbol{h}^\top \boldsymbol{Q} \boldsymbol{h} \ : \ \boldsymbol{h}(\ell) = h(\ell), \ell \in \mathcal{L} \}$.
- Theorem: $\breve{K}(x, x')$ *nearly-linear complexity* in n

Data dependent kernels in nearly-linear time (3)

- Benefit: robustness of using a huge graph (avoid short circuiting), but *efficiently computable*
- state of the art performance on large data-sets in SSL





- also follow ups:
  – efficient CV of many parameters
  – journal version in prep.
  – applying to RL to learn kernels on state space

Tighter PAC-Bayes bounds through distribution dependent priors (1)

- Bounds for stochastic classifiers G_Q drawn from distribution Q on $\mathcal{H}$
- trick is to define PAC-Bayes prior in terms of unknown distribution
- No relative entropy term in bounds
- Exponential weights: density on $\mathcal{H}$ is

$$q(h) = \frac{1}{Z} e^{-\gamma \widehat{\mathrm{risk}}_S(h)} \tag{4}$$

- bound: with probability at least $1 - \delta$,

$$\mathrm{kl}(\widehat{\mathrm{risk}}_S(G_Q), \mathrm{risk}(G_Q)) \leq \frac{1}{m}\left(\gamma \sqrt{\frac{2}{m} \ln \frac{2\sqrt{m}}{\delta}} + \frac{\gamma^2}{2m} + \ln \frac{2\sqrt{m}}{\delta}\right)$$

$\mathrm{kl}(q, p) := q \ln \frac{q}{p} + (1 - q) \ln \frac{1-q}{1-p}$

- no complexity term – only parameter γ

Tighter PAC-Bayes bounds through distribution dependent priors (2)

- RKHS regularization algorithms:

$$h^*_{\mathcal{S}} := \underset{h \in \mathcal{H}_K}{\mathrm{argmin}} \{ \widehat{\mathrm{risk}}^{\ell}_{\mathcal{S}}(h) + \eta ||h||^2_K \} \tag{5}$$

$\mathcal{H}_K$ is RKHS with norm $|| \cdot ||_K$. G_Q is GP with mean and covariance

$$\mathbb{E}[G(x)] = h^*_{\mathcal{S}}(x), \qquad \mathrm{Cov}(G(x), G(x')) = \frac{1}{\gamma} K(x, x') \tag{6}$$

- bound:

$$\mathbb{P}_{\mathcal{S}} \left(\mathrm{kl}(\widehat{\mathrm{risk}}_{\mathcal{S}}(G), \mathrm{risk}(G)) \leq \frac{1}{m} \left(\frac{2\gamma}{\eta^2 m} \ln \frac{8}{\delta} + \ln \frac{4\sqrt{m}}{\delta} \right) \right) \geq 1 - \delta$$

- KL term removed – only parameters η and γ
  – interpreted as complexity terms

Future directions

- We would like to extend the completion to September 2012
- Until September 2012
  1. Extend results on fast online prediction for tree MRFs
  2. Experiments with Bristol data set
- Post September 2012 : Extend p-resistance research
  - UCL and Tuebingen : 2 papers each p-resistance an open research area
  - Visit between UCL and Tuebingen (possibly also Insubria)
  - Some directions:
    1. Computational issues (efficiency + representer theorem)
    2. Loss bounds over the full spectrum of $p \in \infty$
    3. Reinforcement learning application