Active Kernel Learning

Rong Jin

Department of Computer Science and Engineering
Michigan State University, USA

Joint work with Steven H. C. Hoi from Nanyang Technological University

Kernel Learning

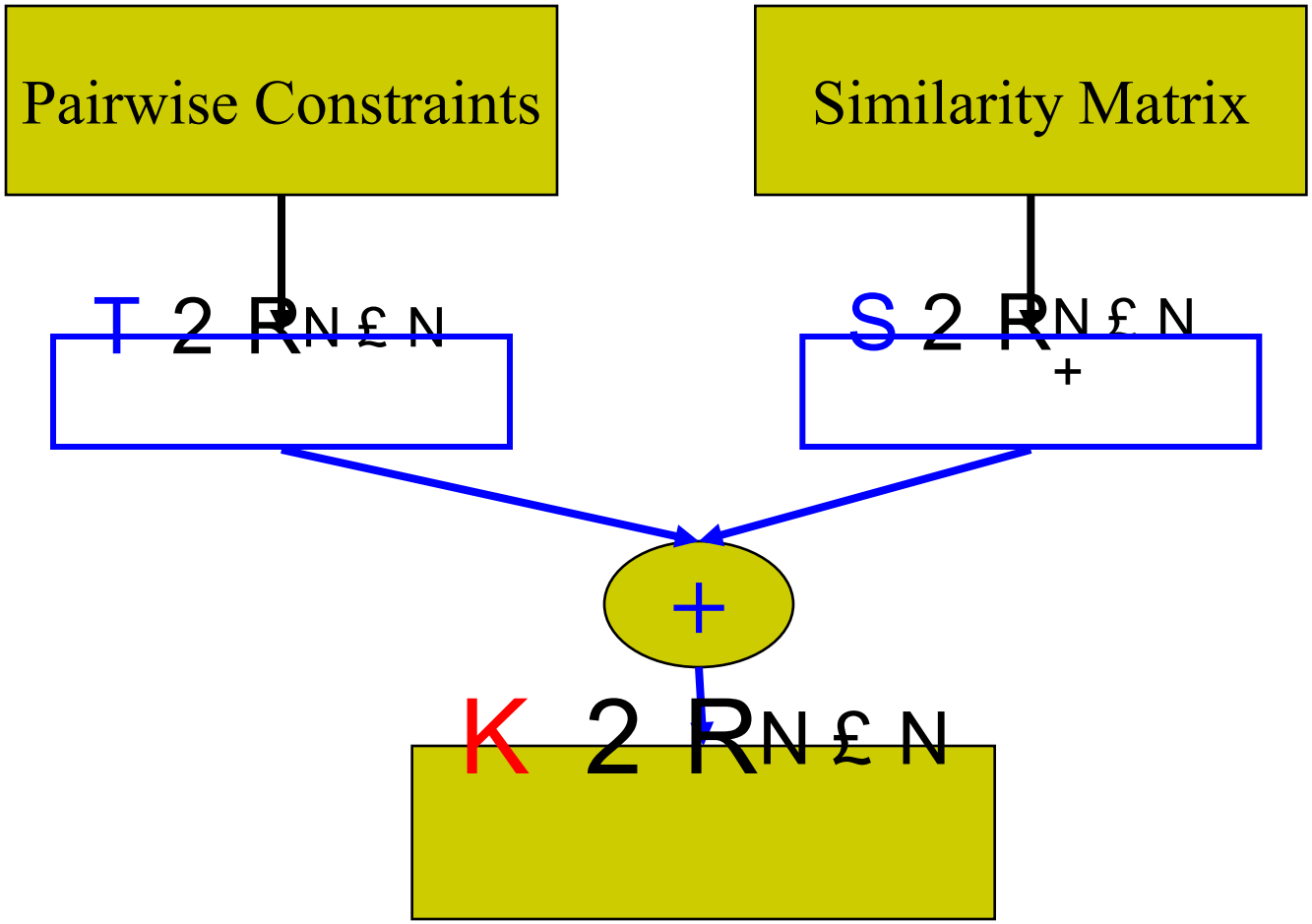
- Kernel is the key to many learning schemes
- Kernel learning
  - Learning parametric kernel functions

  e.g., RBF kernels and diffusion kernels
  - Non-parametric kernel learning
    - No parametric assumption about kernel function
    - Learning kernel matrices

Kernel Learning Methods

- Kernel alignment (Cristianini et al., 2001)
  - Maximize the alignment with the assigned class labels
- Learning kernel with SDP (Lanckriet et al., 2004)
  - Maximize the classification margin
- Nonparametric graph kernel (Zhu et al., 2005)
  - Kernel alignment + order constraint
- Learning low rank kernel matrices (Kulis et al., 2006)
  - Initial kernel + consistent with the assigned class labels
- Non-parametric kernel learning (Hoi et. al., 2007)
  - Initial similarity + pairwise constraints

Non-parametric Kernel Learning



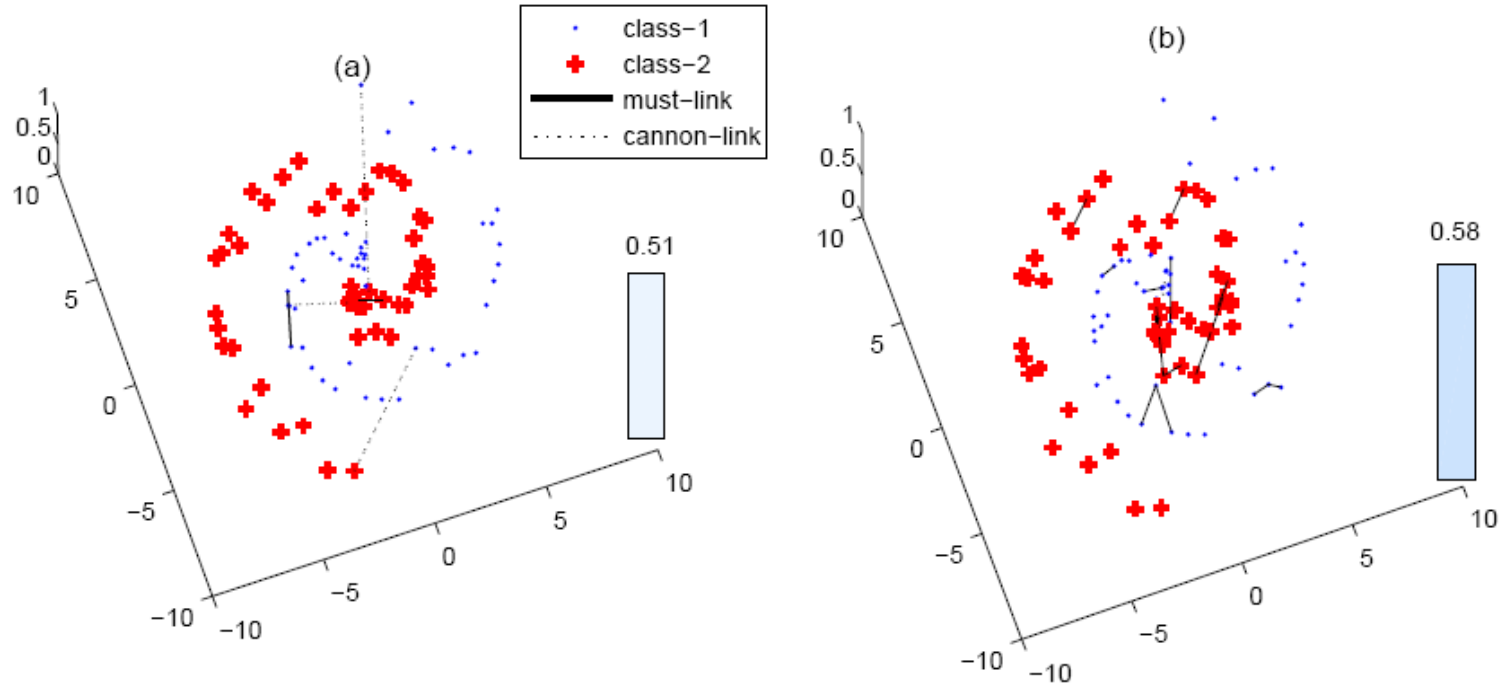
Active Kernel Learning

- Existing kernel learning schemes are passive
  - Assume labeling information is given beforehand
  - Less efficient, may require more labeling efforts
- Active kernel learning
  - Select the most informative example pairs from a pool of unlabeled examples
  - Extension of non-parametric kernel learning (Hoi et. al., 2007)

Active Kernel Learning

- Key challenge
  - How to measure the informativeness for an example pair ?
- Simple strategy: uncertainty principle
  - Select an example pair with the least confidence
    - Example pairs with least $\|K_{i,j}\|$
  - Main drawback
    - Lacks of principles
    - Could result in the preference of must-link constraints

Active Kernel Learning



Uncertainty principle leads to the preference of must-link constraints

Contribution

- Propose a min-max framework for active kernel learning
- Present a convex relaxation for the proposed framework of active kernel learning

Non-parametric Kernel Learning: Review

- Data $U = (x_1; x_2; \cdots; x_N); \quad x \in R^d$

- **Pairwise constraints**: $T \in R^{N \times N}$

$$T_{i;j} = \begin{cases} 1 & x_i \text{ and } x_j \text{ in the same class} \\ -1 & x_i \text{ and } x_j \text{ in different classes} \\ 0 & \text{otherwise} \end{cases}$$

must-link

cannot-link

- **Similarity measurements**: $S \in R_+^{N \times N}$

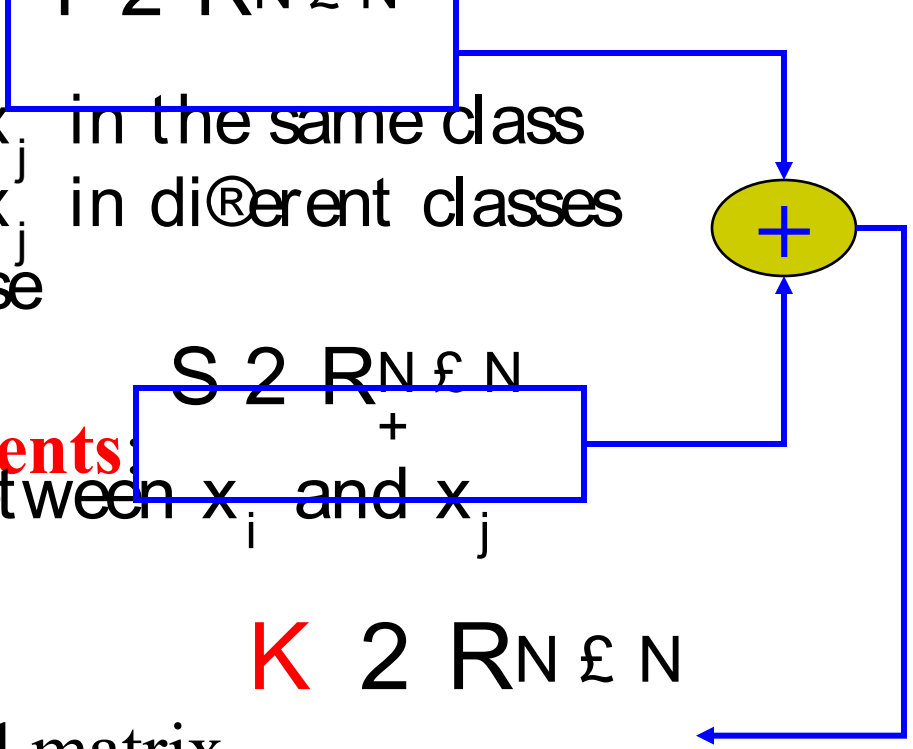
  - $S_{i;j}$: similarity between x_i and x_j

Non-parametric Kernel Learning: Review

- Data $U = (x_1; x_2; \cdots; x_N); \quad x \in R^d$
- **Pairwise constraints**: $T \in R^{N \times N}$

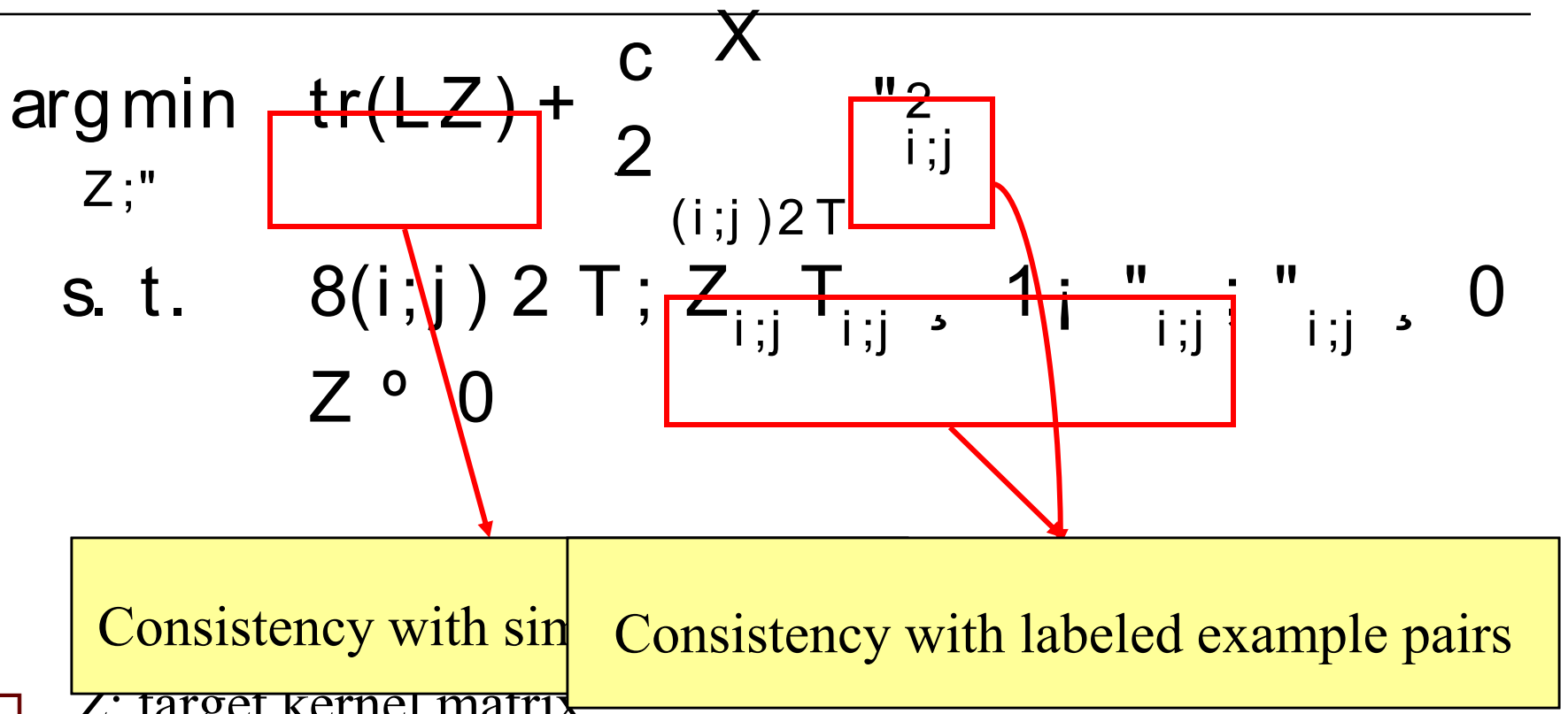
$$T_{i;j} = \begin{cases} 1 & x_i \text{ and } x_j \text{ in the same class} \\ -1 & x_i \text{ and } x_j \text{ in different classes} \\ 0 & \text{otherwise} \end{cases}$$

- **Similarity measurements**: $S \in R^{N \times N}_{+}$
  - $S_{i;j}$: similarity between x_i and x_j



Learning a kernel matrix

Non-parametric Kernel Learning



- Z: target kernel matrix
- L: graph Laplacian for similarity matrix S

Min-Max Framework

- Informativeness of an unlabeled pair (x_k, x_l)

$$!((k,l),y) = \min_{Z,\xi} \; \mathrm{tr}(LZ) + \frac{c}{2}\sum_{(i,j)\in T} \xi_{ij}^2 + \frac{c}{2}\xi_{kl}^2$$

$$\text{s.t.} \; T_{ij} Z_{ij} \geq 1 - \xi_{ij}, \; \forall (i,j) \in T$$

$$y Z_{kl} \geq 1 - \xi_{kl}, \; Z \succeq 0$$

Measures how example pair (k,l) with label y will affect the overall objective function.

Min-Max Framework

- Informativeness of an unlabeled pair $(x_k; x_l)$

$$
\begin{aligned}
!((k;l);y) = \quad & \min_{Z,\xi} \quad \mathrm{tr}(LZ) + \frac{c}{2}\sum_{(i,j)\in T}\xi_{i,j}^2 + \frac{c}{2}\xi_{k,l}^2 \\
& \text{s.t.} \quad T_{i,j}Z_{i,j} \geq 1-\xi_{i,j}, \ \forall (i,j)\in T \\
& \qquad yZ_{k,l} \geq 1-\xi_{k,l}, \ Z \succeq 0
\end{aligned}
$$

(k,l) is highly consistent with the current kernel Z

$!((k;l);y)$ ⇩ will be small with appropriate choice of y,

and will be large with incorrect y

Min-Max Framework

- Informativeness of an unlabeled pair $(x_k; x_l)$

$$! ((k;l);y) = \min_{Z;"} \; tr(LZ) + \frac{c}{2} \sum_{(i;j)2T} "^2_{i;j} + \frac{c}{2} "^2_{k;l}$$

$$\text{s. t. } T_{\cdot} . Z_{\cdot} . _{\cdot} . \; 1_{\dot{\iota}} \; "_{\cdot} . _{\cdot} ; \; 8(i;j) 2 T$$

Measure the uninformativeness of an example pair $\cdot (k;l)$

example pairs highly consistent with Z → large

$$\cdot (k;l) = \max_{y2 f \dot{\iota} \; 1;+ \; 1g} \; ! ((k;l);y)$$

Min-Max Framework

$$(k;l)^{\alpha} = \arg\min_{(k;l)\in T} \cdot (k;l) = \arg\min_{(k;l)\in T} \max_{t\in\{j\ 1;+\ 1g} ! ((k;l);t) \qquad (P)$$

- A challenging optimization problem
  - $! ((k;l);t)$ is not a closed-form function
  - Min-max optimization

Min-Max Framework: Simplification

Theorem 1: (P) is equivalent to

$$\min_{Z,\xi,(k,l)\in T} \quad \mathrm{tr}(LZ) + c\sum_{(i,j)\in T}\xi_{ij}^2 + c\xi_{kl}^2$$

$$\text{s.t.} \quad T_{ij}Z_{ij} \geq 1-\xi_{ij},\ \xi_{ij}\geq 0,\ \forall (i,j)\in T \qquad (P1)$$

$$\xi_{kl}\geq 1+|Z_{kl}|,\ Z\succeq 0$$

Corollary 2: When Z is fixed, ($P1$) is equivalent to

$$(k,l)^* = \arg\min_{(k,l)\in T} |Z_{kl}|$$

Min-Max Framework: Simplification

Theorem 1: (P) is equivalent to

$$\min_{Z,\xi,(k,l)\in T} \; \mathrm{tr}(LZ) + c\sum_{(i,j)\in T}\xi_{ij}^2 + c\xi_{kl}^2$$

$$\text{s.t.} \quad T_{ij}Z_{ij} \geq 1-\xi_{ij}, \; \xi_{ij} \geq 0, \; \forall (i,j)\in T \qquad (P1)$$

$$\xi_{kl} \geq 1 + |Z_{kl}|, \; Z \succeq 0$$

1. Introduce p_{kl} to weight each example pair (k,l)
2. Aggregate all the optimization problems together

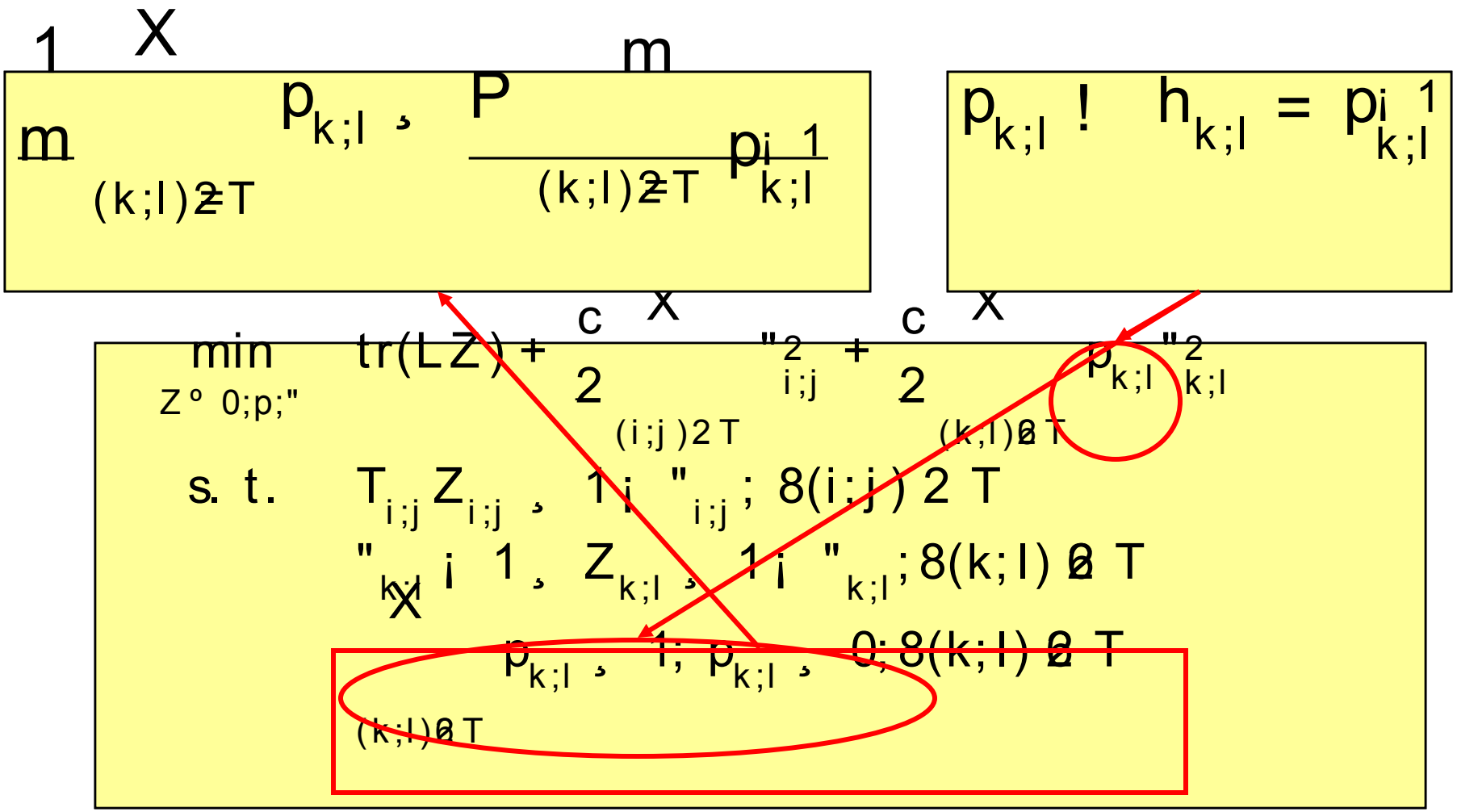
Min-Max Framework: Simplification

Theorem 3: ($P1$) is equivalent to

$$
\begin{aligned}
\min_{Z \succeq 0;\, \rho;\, \varepsilon} \quad & \operatorname{tr}(LZ) + \frac{c}{2} \sum_{(i,j) \in T} \varepsilon_{i,j}^2 + \frac{c}{2} \sum_{(k,l) \notin T} \rho_{k,l}\, \varepsilon_{k,l}^2 \\
\text{s.t.} \quad & T_{i,j} - Z_{i,j} \le 1 - \varepsilon_{i,j}, \; \forall (i,j) \in T \\
& \varepsilon_{k,l} - 1 \le Z_{k,l} \le 1 - \varepsilon_{k,l}, \; \forall (k,l) \notin T \\
& \sum_{(k,l) \notin T} \rho_{k,l} \le 1, \; \rho_{k,l} \ge 0, \; \forall (k,l) \notin T
\end{aligned}
\tag{$P2$}
$$

But, this is non-convex optimization !

Min-Max Framework: Simplification



Min-Max Framework: Simplification

$$\min_{Z \succeq 0;\, \eta_{k,l}} \; \mathrm{tr}(LZ) + \frac{c}{2} \sum_{(i,j) \in T} \xi_{i,j}^2 + \frac{c}{2} \sum_{(k,l) \notin T} \eta_{k,l}^2$$

$$\text{s.t.} \quad T_{i,j} Z_{i,j} \geq 1 - \xi_{i,j}; \; \xi_{i,j} \geq 0, \; \forall (i,j) \in T$$

$$\eta_{k,l} \geq 1 - Z_{k,l}, \; \eta_{k,l} \geq 0, \; \forall (k,l) \notin T$$

$$\sum_{(k,l) \notin T} \eta_{k,l} \leq m^2; \; \eta_{k,l} \leq 1, \; \forall (k,l) \notin T$$

$$\eta_{k,l} = p_{k,l}^{i-1}$$

Theorem 4:

- ($P3$) is convex (semi-definite programming)
- The optimal value of ($P2$) is upper bounded by that of ($P3$)

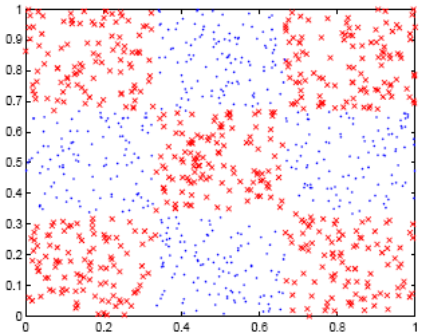
Min-Max Framework: Simplification

$$
\begin{aligned}
\min_{Z \succeq 0,\, h,\, \varepsilon} \quad & \operatorname{tr}(LZ) + \frac{c}{2} \sum_{(i,j) \in T} \varepsilon_{i,j}^2 + \frac{c}{2} \sum_{(k,l) \notin T} \varepsilon_{k,l}^2 \\
\text{s.t.} \quad & T_{i,j} Z_{i,j} \geq 1 - \varepsilon_{i,j}, \ \varepsilon_{i,j} \geq 0, \ \forall (i,j) \in T \\
& \varepsilon_{k,l} \geq 1 - Z_{k,l} \geq 1 - \varepsilon_{k,l}, \ \forall (k,l) \notin T \\
& \sum_{(k,l) \notin T} h_{k,l} \cdot m^2, \ h_{k,l} \geq 1, \ \forall (k,l) \notin T
\end{aligned}
\qquad (P3)
$$

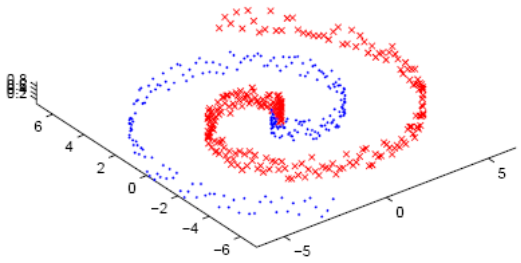
($P3$) can be optimized more efficiently using the dual formulation (details in paper)

Experiments: Datasets

Dataset	#Classes	#Instances	#Features
Chessboard	2	100	2
Double-Spiral	2	100	3
Glass	6	214	9
Heart	2	270	13
Iris	3	150	4
Protein	6	116	20
Sonar	2	208	60
Soybean	4	47	35
Wine	3	178	12



Chessboard

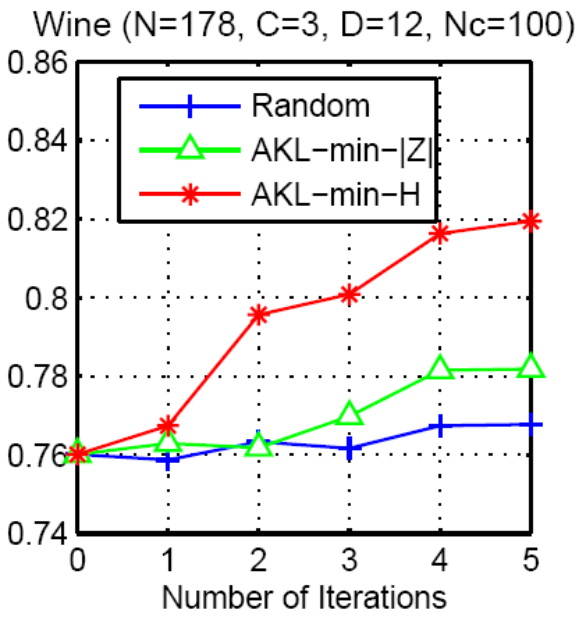


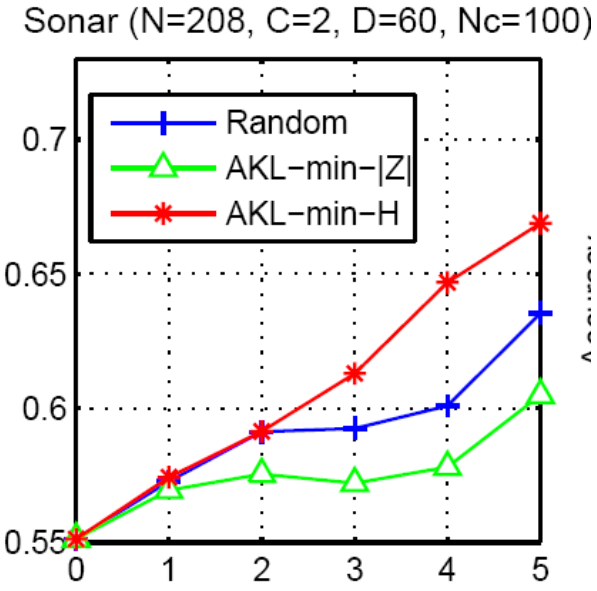
Double Spiral

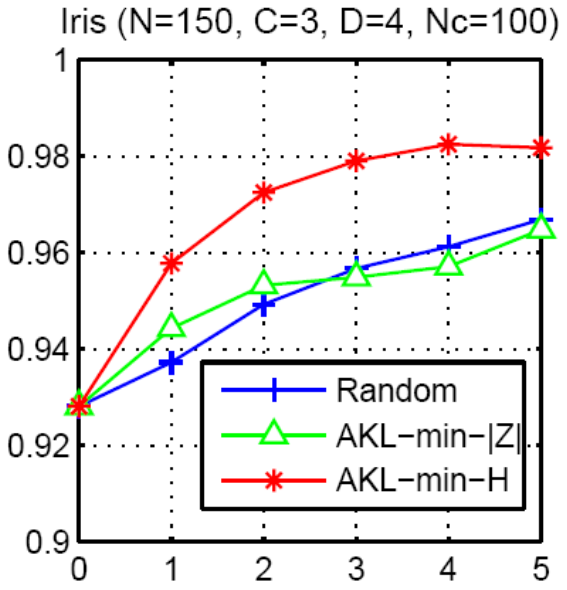
Experiments: Setup

- Baselines
  - **Random**: randomly samples example pairs
  - **AKL-min-|Z|**: chooses pair examples with the least $|Z_{k,l}|$
  - **AKL-min-H**: the proposed AKL algorithm. selects example pairs with least $h_{k,l}$ that corresponds to maximal $p_{k,l}$.
- Evaluation
  - Evaluate kernel matrices by data clustering using kernel K-means
  - Clustering accuracy = $\sum_{i>j} \frac{1\{1\{c_i = c_j\} = 1\{\hat{c}_i = \hat{c}_j\}\}}{0.5n(n-1)}$

Experimental Results







1. Nc: number of initially given pairwise constraints
2. 20 additional pairiwise constraints are provided each iteration

AKL-min-H > Random ≈ AKL-min-|Z|

Conclusion

- Propose a min-max framework for active learning, and apply it to active kernel learning
- Present a convex relaxation for the proposed min-max framework for active kernel learning
- Show encouraging experimental results using data clustering