# Learning Kernels

Mehryar Mohri
Courant Institute and Google Research
mohri@cs.nyu.edu

Joint work with
Corinna Cortes and Afshin Rostamizadeh

# Motivation

- Learning kernels:
  - appealing goal and active area of research.
  - lots of questions!
- Workshop opportunity:
  - brief survey of latest theoretical results.
  - brief survey of techniques and empirical results.

# This Talk

- Theoretical results
- Techniques and empirical results

# Hypothesis Set

- Convex combination of $p$ base kernels:

$$\mathcal{K} = \Big\{ \sum_{k=1}^{p} \mu_k K_k \colon \mu_k \geq 0, \sum_{k=1}^{p} \mu_k = 1 \Big\}.$$

- Hypothesis set [Lanckriet et al., 2004]:

$$H = \Big\{ x \mapsto \sum_{i=1}^{m} \alpha_i K(x_i, x) \colon K \in \mathcal{K}, \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \leq 1/\rho^2 \Big\}.$$

# Previous Learning Bounds

- For any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H$, [Bousquet and Herrmann 2003; Lanckriet et al., 2004]

$$R(h) \leq \widehat{R}_\rho(h) + \frac{1}{\sqrt{m}} \left[ \sqrt{\frac{\max_{k=1}^p \operatorname{Tr}(\mathbf{K}_k) \max_{i=1}^p \frac{\|\mathbf{K}_k\|}{\operatorname{Tr}(\mathbf{K}_k)}}{\rho^2} + 4 + \sqrt{2 \log \frac{1}{\delta}} \right].$$

- But, bound always greater than one! [Srebro and Ben-David, 2006]

- Other bound for linear combination case ([Lanckriet et al., 2004]) also always greater than one.

# Multiplicative Learning Bound

[Lanckriet et al., 2004]

- For any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H$,

$$R(h) \leq \widehat{R}_\rho(h) + O\left(\sqrt{\frac{p/\rho^2}{m}}\right).$$

- Bound multiplicative in $p$ (number of kernels).

# Additive Learning Bound

[Srebro and Ben-David, 2006]

- Assume that for all $k \in [1, p]$, $K_k(x, x) \leq R^2$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H$,

$$R(h) \leq \widehat{R}_\rho(h) + \sqrt{8 \frac{2 + p \log \frac{128 e m^3 R^2}{\rho^2 p} + 256 \frac{R^2}{\rho^2} \log \frac{\rho e m}{8R} \log \frac{128 m R^2}{\rho^2} + \log(1/\delta)}{m}}.$$

- Bound additive in $p$ (modulo log terms).
  - based on pseudo-dimension of kernel family.
  - similar guarantees for other families.

# New Learning Bound

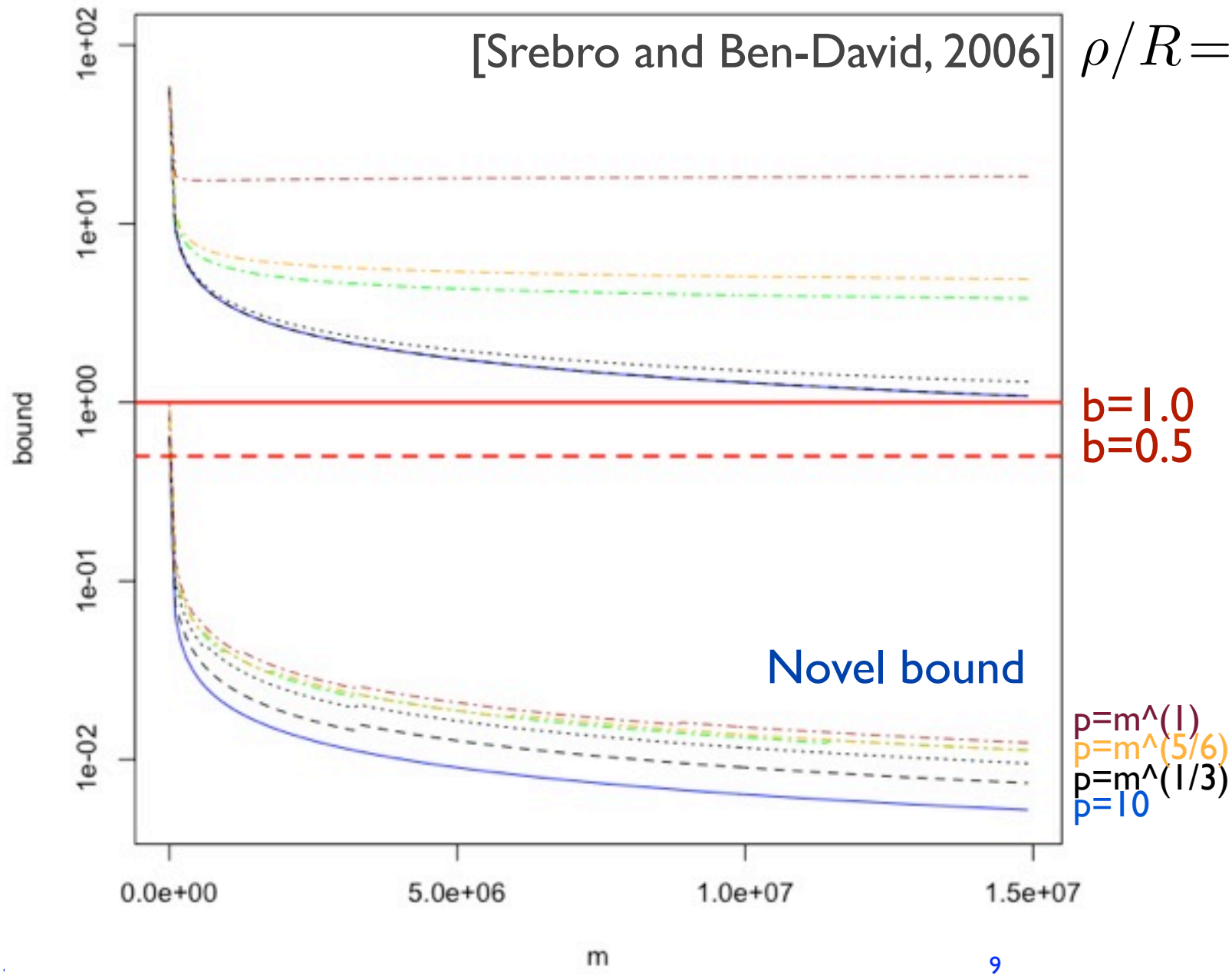
[Cortes, MM, and Rostami, (2010)]

- Assume that for all $k \in [1, p]$, $K_k(x, x) \leq R^2$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H$,

$$R(h) \leq \widehat{R}_\rho(h) + O\left(\sqrt{\frac{(\log p) R^2/\rho^2}{m}}\right).$$

- Very weak dependence on $p$, no extra log terms:
  - analysis based on Rademacher complexity.
  - bound valid for $p \gg m$ .

# Plots


[Srebro and Ben-David, 2006]
$\rho/R = .2$
bound
1e+02
1e+01
1e+00
1e-01
1e-02
b=1.0
b=0.5
Novel bound
p=m^(1)
p=m^(5/6)
p=m^(1/3)
p=10
0.0e+00
5.0e+06
1.0e+07
1.5e+07
m

# New Learning Bound - L2 Reg.

[Cortes, MM, and Rostami, (2010)]

- Assume that for all $k \in [1, p]$, $K_k(x, x) \leq R^2$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H$,

$$R(h) \leq \widehat{R}_\rho(h) + O\left(p^{1/4}\sqrt{\frac{R^2/\rho^2}{m}}\right).$$

- Mild dependence on $p$, no log terms.

# Learning Bound - L2 Regularization

[Cortes, MM, and Rostami, 2009]

- Regression, KRR $L_2$ regularization, stability bound:

$$R(h) \leq \widehat{R}(h) + O\left(\sqrt{p/m} + \sqrt{1/m}\right)$$

  - additive term with number of kernels $p$.
  - technical condition (orthogonal kernels).
  - suggests using larger number of kernels $p$.

# This Talk

- Theoretical results
- Techniques and empirical results

# Brief Survey

- Early learning kernel work.
- Standard optimization problem.
- L2 regularization.
- Non-linear combinations.

# Learning Kernels

- Standard SVM dual optimization problem:

$$\max_{\boldsymbol{\alpha}} \quad 2\boldsymbol{\alpha}^\top \mathbf{1} - \boldsymbol{\alpha}^\top \mathbf{Y}^\top \mathbf{K} \mathbf{Y} \boldsymbol{\alpha}$$

$$\text{subject to} \quad \boldsymbol{\alpha}^\top \mathbf{y} = 0 \ \wedge \ \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{C}$$

Structural Risk Minimization: select the kernel that minimizes an estimate of the generalization error.

- What estimate should we minimize?

# Minimize Different Criterion

(Chapelle, Vapnik, Bousquet & Mukherjee, 2000)

- Alternate SVM and gradient step algorithm:
  - solve SVM problem and obtain $\alpha^\star$.
  - gradient step over criterion $T$ to select kernel parameters:
    - margin criterion $T = R^2/\rho^2$.
    - span criterion $T = \frac{1}{m}\sum_{i=1}^{m} \Theta(\alpha_i^\star S_i^2 - 1)$.
- Other similar work:
  - margin criterion $T = R^2/\rho^2$ (Weston et al., NIPS 2001).

# Empirical Results

(Chapelle, Vapnik, Bousquet & Mukherjee, 2000)

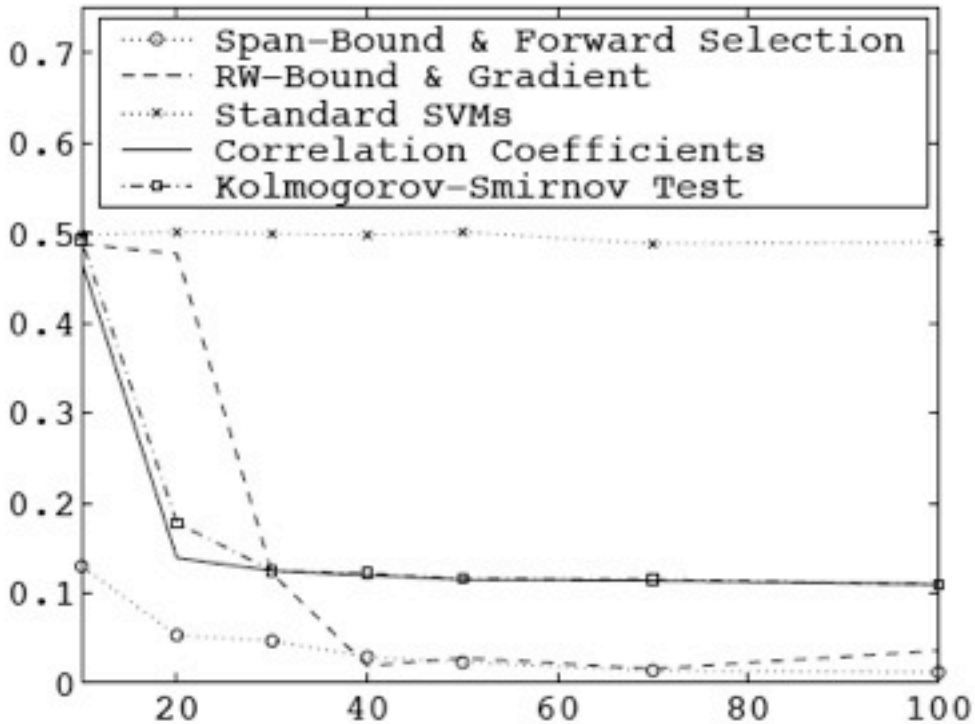
| | Cross-validation | $R^2/M^2$ | Span-bound |
|---|---|---|---|
| Breast Cancer | $26.04 \pm 4.74$ | $26.84 \pm 4.71$ | $25.59 \pm 4.18$ |
| Diabetis | $23.53 \pm 1.73$ | $23.25 \pm 1.7$ | $23.19 \pm 1.67$ |
| Heart | $15.95 \pm 3.26$ | $15.92 \pm 3.18$ | $16.13 \pm 3.11$ |
| Thyroid | $4.80 \pm 2.19$ | $4.62 \pm 2.03$ | $4.56 \pm 1.97$ |
| Titanic | $22.42 \pm 1.02$ | $22.88 \pm 1.23$ | $22.5 \pm 0.88$ |

- Selecting the width of a Gaussian kernel and the SVM parameter C.
- 100x faster than cross-validation.

# Experiments - Feature Selection

(Chapelle, Vapnik, Bousquet & Mukherjee, 2000; Weston et al., NIPS 2001)

- Comparison with existing methods:



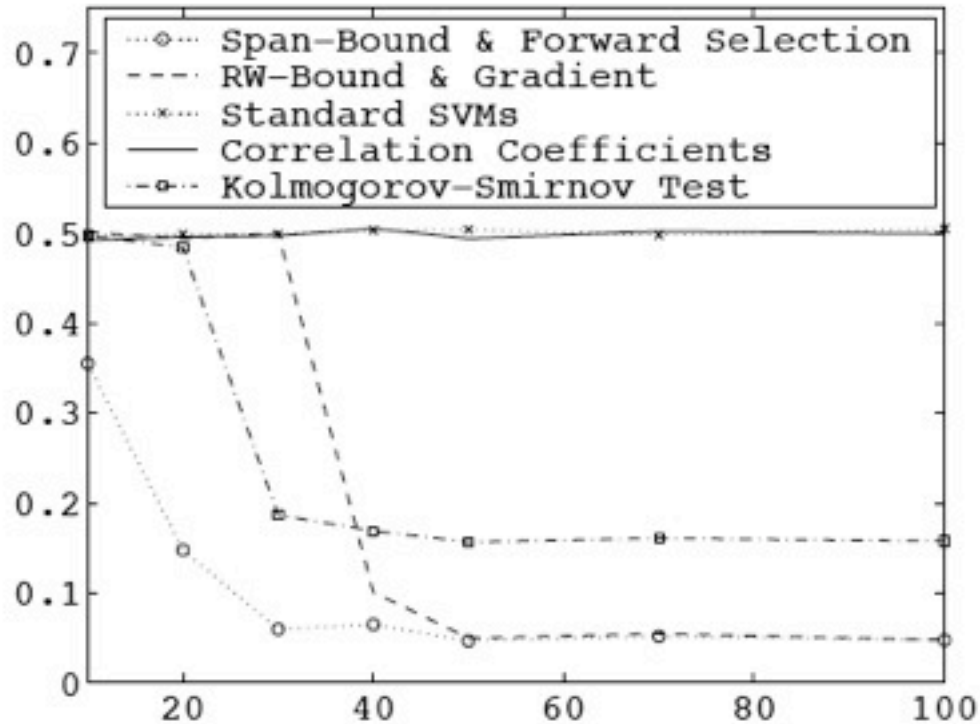




Figure 1: A comparison of feature selection methods on (a) a linear problem and (b) a nonlinear problem both with many irrelevant features. The $x$-axis is the number of training points, and the $y$-axis the test error as a fraction of test points.

# SVM - Learning Kernels Formulation

(Lanckriet et al., 2003)

see also (Grandvalet & Canu, 2002)

- Optimization problem:

$$\min_{\boldsymbol{\mu}} \max_{\boldsymbol{\alpha}} \; F(\boldsymbol{\mu}, \boldsymbol{\alpha}) = 2\boldsymbol{\alpha}^\top \mathbf{1} - \boldsymbol{\alpha}^\top \mathbf{Y}^\top \left( \sum_{k=1}^{p} \mu_k \mathbf{K}_k \right) \mathbf{Y} \boldsymbol{\alpha}$$

$$\text{subject to} \quad \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{C} \wedge \boldsymbol{\alpha}^\top \mathbf{y} = 0$$

$$\boldsymbol{\mu} \geq \mathbf{0} \wedge \sum_{k=1}^{p} \mu_k \mathrm{Tr}(\mathbf{K}_k) \leq \Lambda.$$

- Solution: SDP in general, QCQP or SILP for linear combination, or QP for rank-1.

# Empirical Results

(Lanckriet et al., 2003)

| | | $K_1$ | $K_2$ | $K_3$ | $\sum_i \mu_i^* K_i$ | $\sum_i \mu_{i,+}^* K_i$ | best c/v RBF |
|---|---|---|---|---|---|---|---|
| *Heart* | | $d = 2$ | $\sigma = 0.5$ | | | | |
| HM | $\gamma$ | 0.0369 | 0.1221 | - | 0.1531 | 0.1528 | |
| | **TSA** | **72.9 %** | **59.5 %** | - | **84.8 %** | **84.6 %** | **77.7 %** |
| | $\mu_1/\mu_2/\mu_3$ | 3/0/0 | 0/3/0 | 0/0/3 | **-0.09/2.68/0.41** | **0.01/2.60/0.39** | |
| SM1 | $\omega_{S1}^*$ | 58.169 | 33.536 | 74.302 | 21.361 | 21.446 | |
| | **TSA** | **79.3 %** | **59.5 %** | **84.3 %** | **84.8 %** | **84.6 %** | **83.9 %** |
| | $C$ | 1 | 1 | 1 | 1 | 1 | |
| | $\mu_1/\mu_2/\mu_3$ | 3/0/0 | 0/3/0 | 0/0/3 | **-0.09/2.68/0.41** | **0.01/2.60/0.39** | |
| SM2 | $\omega_{S2}^*$ | 32.726 | 25.386 | 45.891 | 15.988 | 16.034 | |
| | **TSA** | **78.1 %** | **59.0 %** | **84.3 %** | **84.8 %** | **84.6 %** | **83.2 %** |
| | $C$ | 1 | 1 | 1 | 1 | 1 | |
| | $\mu_1/\mu_2/\mu_3$ | 3/0/0 | 0/3/0 | 0/0/3 | **-0.08/2.54/0.54** | **0.01/2.47/0.53** | |
| SM2,C | $\omega_{S2}^*$ | 19.643 | 25.153 | 16.004 | | 15.985 | |
| | **TSA** | **81.3 %** | **59.6 %** | **84.7 %** | | **84.6 %** | **83.2 %** |
| | $C$ | 0.3378 | 1.18e+7 | 0.2880 | | 0.4365 | |
| | $\mu_1/\mu_2/\mu_3$ | 1.04/0/0 | 0/3.99/0 | 0/0/0.53 | | **0.01/0.80/0.53** | |

# Empirical Results

(Lanckriet, De Bie, Cristianini, Jordan, & Noble, 2004)

- Classification - cytoplasmic ribosomal class.

  Performance measured wrt ranking criterion (AUC).

| $K_{SW}$ | $K_{PF}$ | $K_{LI}$ | $K_B$ | $K_D$ | $K_{R1...R6}$ | $K_{R7...R12}$ | TP1FP | ROC |
|---|---|---|---|---|---|---|---|---|
| 5.08 | 0.31 | 0.22 | 0.39 | 0.00 | – | – | $88.21 \pm 1.73\%$ | $0.9933 \pm 0.0011$ |
| 5.07 | 0.31 | 0.22 | 0.39 | 0.00 | 0.01 | – | $88.19 \pm 1.60\%$ | $0.9932 \pm 0.0011$ |
| 5.06 | 0.30 | 0.22 | 0.38 | 0.01 | 0.02 | 0.01 | $88.08 \pm 1.65\%$ | $0.9932 \pm 0.0010$ |
| 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | – | – | $75.20 \pm 2.38\%$ | $0.9906 \pm 0.0012$ |
| 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | – | $59.66 \pm 3.03\%$ | $0.9791 \pm 0.0017$ |
| 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | $42.87 \pm 2.59\%$ | $0.9620 \pm 0.0027$ |

# Hyperkernels

(Ong, Smola & Williamson, 2005)

- Kernels of kernels, infinitely many kernels.
- $m^2$ kernel parameters to optimize over.

$$K(x, x') = \sum_{i,j=1}^{m} \beta_{i,j} \underline{K}((x_i, x_j), (x, x'))$$

$$\forall x, x' \in X, \quad \beta_{i,j} \geq 0$$

- SDP problem.

# Empirical Results

| Data | $C$-SVM | $\nu$-SVM | Lag-SVM | Best other | CV Tuned SVM ($C$) |
|---|---|---|---|---|---|
| syndata | 2.8±2.4 | **1.9±1.9** | 2.4±2.2 | NA | 5.9±5.4 ($10^8$) |
| pima | **23.5±2.0** | 27.7±2.1 | 23.6±1.9 | 23.5 | 24.1±2.1 ($10^4$) |
| ionosph | 6.6±1.8 | 6.7±1.8 | 6.4±1.9 | **5.8** | 6.1±1.8 ($10^3$) |
| wdbc | 3.3±1.2 | 3.8±1.2 | **3.0±1.1** | 3.2 | 5.2±1.4 ($10^6$) |
| heart | 19.7±3.3 | 19.3±2.4 | 20.1±2.8 | **16.0** | 23.2±3.7 ($10^4$) |
| thyroid | 7.2±3.2 | 10.1±4.0 | 6.2±3.1 | **4.4** | 5.2±2.2 ($10^5$) |
| sonar | 14.8±3.7 | 15.3±3.7 | **14.7±3.6** | 15.4 | 15.3±4.1 ($10^3$) |
| credit | 14.6±1.8 | **13.7±1.5** | 14.7±1.8 | 22.8 | 15.3±2.0 ($10^8$) |
| glass | 6.0±2.4 | 8.9±2.6 | **6.0±2.2** | NA | 7.2±2.7 ($10^3$) |

$$\underline{K}\Big((x,x'),(x'',x''')\Big) = \prod_{j=1}^{d} \frac{1-\lambda}{1-\lambda\exp\Big(-\sigma_j\big((x_j-x'_j)^2+(x''_j-x'''_j)^2\big)\Big)}$$

# Regression - L2 Regularization

[Cortes, MM, and Rostami, 2009]

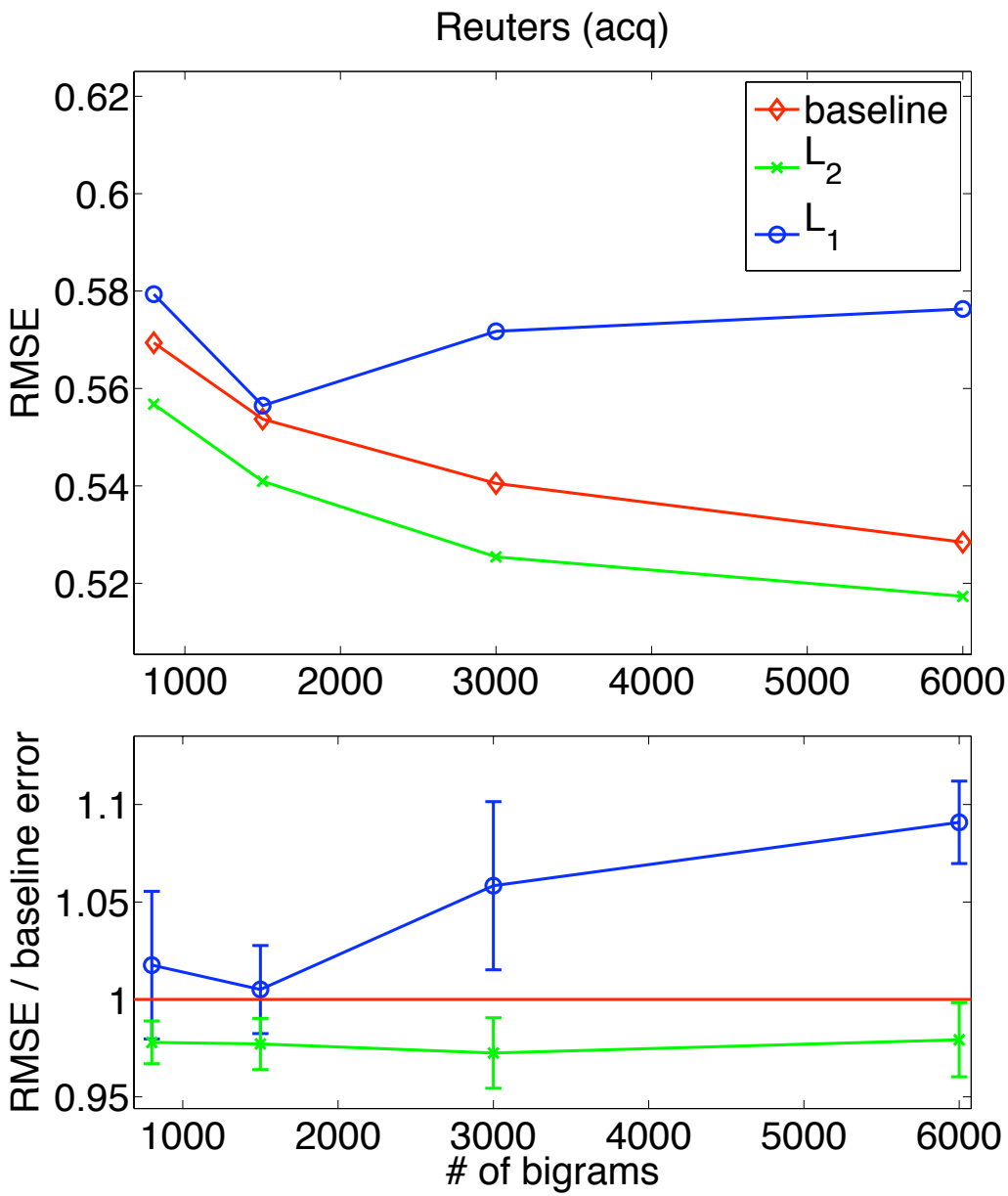
- Optimization problem:

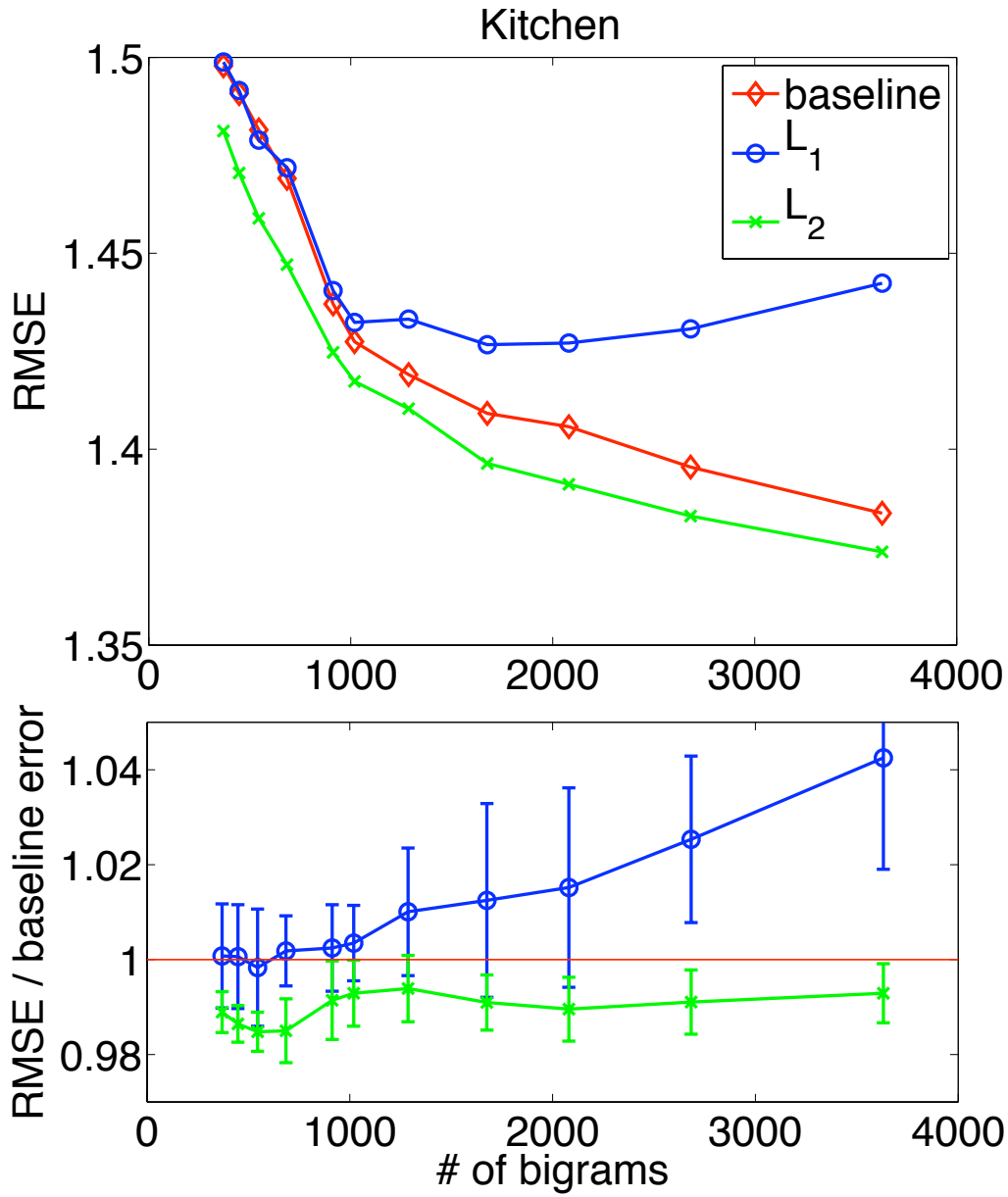
$$\min_{\boldsymbol{\mu} \in \mathcal{M}} \max_{\boldsymbol{\alpha}} -\lambda \boldsymbol{\alpha}^\top \boldsymbol{\alpha} - \sum_{k=1}^{p} \mu_k \boldsymbol{\alpha}^\top \mathbf{K}_k \boldsymbol{\alpha} + 2\boldsymbol{\alpha}^\top \mathbf{y}$$

with $\mathcal{M} = \{\boldsymbol{\mu} \colon \boldsymbol{\mu} \geq 0 \wedge \|\boldsymbol{\mu} - \boldsymbol{\mu}_0\|^2 \leq \Lambda^2\}$.

- Solution:
  - simple iterative algorithm.
  - very efficient in practice.
- Results: L2 regularization leads to better results (see also [Kloft et al., 2009, this conference]).

# Empirical Results - Rank-1 Kernels



# Non-Linear Combinations

- DC-Programming algorithm (Argyriou et al., 2005)
- Generalized MKL (Varma & Babu, 2009)
- Hierarchical kernels (Bach, 2008)
- Polynomial combinations (Cortes, MM, and Rostami, 2009)

# Hierarchical Kernel Learning

(Bach, 2008)

- Example:
  - Sub kernel:

$$K_{i,j}(x_i, x_i') = \binom{q}{j}(1 + x_i x_i')^j, \quad i \in [1, p], \quad j \in [0, q]$$

  - Full kernel:

$$K(x, x') = \prod_{i=1}^{p}(1 + x_i x_i')^q$$

  - Convex optimization problem, complexity polynomial in the number of kernels selected, sparsity through $L_1$ regularization and hierarchical selection criteria.

# Reality Check - HKL

| dataset | $n$ | $p$ | $k$ | $\#(V)$ | L2 | MKL | HKL |
|---|---|---|---|---|---|---|---|
| abalone | 4177 | 10 | pol4 | $\approx 10^{7}$ | 44.2±1.3 | 44.5±1.1 | **43.3±1.0** |
| abalone | 4177 | 10 | rbf | $\approx 10^{10}$ | **43.0±0.9** | 43.7±1.0 | 43.0±1.1 |
| bank-32fh | 8192 | 32 | pol4 | $\approx 10^{22}$ | 40.1±0.7 | **38.7±0.7** | 38.9±0.7 |
| bank-32fh | 8192 | 32 | rbf | $\approx 10^{31}$ | 39.0±0.7 | 38.4±0.7 | **38.4±0.7** |
| bank-32fm | 8192 | 32 | pol4 | $\approx 10^{22}$ | 6.0±0.1 | 6.1±0.3 | 5.1±0.1 |
| bank-32fm | 8192 | 32 | rbf | $\approx 10^{31}$ | 5.7±0.2 | 5.9±0.2 | **4.6±0.2** |
| bank-32nh | 8192 | 32 | pol4 | $\approx 10^{22}$ | 44.3±1.2 | 46.0±1.2 | **43.6±1.1** |
| bank-32nh | 8192 | 32 | rbf | $\approx 10^{31}$ | 44.3±1.2 | 46.1±1.1 | **43.5±1.0** |
| bank-32nm | 8192 | 32 | pol4 | $\approx 10^{22}$ | 17.2±0.6 | 21.0±0.7 | **16.8±0.6** |
| bank-32nm | 8192 | 32 | rbf | $\approx 10^{31}$ | 16.9±0.6 | 20.9±0.7 | **16.4±0.6** |
| boston | 506 | 13 | pol4 | $\approx 10^{9}$ | **17.1±3.6** | 22.2±2.2 | 18.1±3.8 |
| boston | 506 | 13 | rbf | $\approx 10^{12}$ | **16.4±4.0** | 20.7±2.1 | 17.1±4.7 |
| pumadyn-32fh | 8192 | 32 | pol4 | $\approx 10^{22}$ | 57.3±0.7 | **56.4±0.7** | 56.4±0.8 |
| pumadyn-32fh | 8192 | 32 | rbf | $\approx 10^{31}$ | 57.7±0.6 | 56.5±0.8 | **55.7±0.7** |
| pumadyn-32fm | 8192 | 32 | pol4 | $\approx 10^{22}$ | 6.9±0.1 | 7.0±0.1 | **3.1±0.0** |
| pumadyn-32fm | 8192 | 32 | rbf | $\approx 10^{31}$ | 5.0±0.1 | 7.1±0.1 | **3.4±0.0** |
| pumadyn-32nh | 8192 | 32 | pol4 | $\approx 10^{22}$ | 84.2±1.3 | 83.6±1.3 | **36.7±0.4** |
| pumadyn-32nh | 8192 | 32 | rbf | $\approx 10^{31}$ | 56.5±1.1 | 83.7±1.3 | **35.5±0.5** |
| pumadyn-32nm | 8192 | 32 | pol4 | $\approx 10^{22}$ | 60.1±1.9 | 77.5±0.9 | **5.5±0.1** |
| pumadyn-32nm | 8192 | 32 | rbf | $\approx 10^{31}$ | 15.7±0.4 | 77.6±0.9 | **7.2±0.1** |

# Conclusion

- Theory: generalization bounds mildly dependent on $p$.
- Empirical results:
  - not consistently significantly > even combination.
  - L2 regularization seems to help.
  - large number of kernels seems to help.
  - non-linear combination seems to help.
  - other benefits: feature selection, speed, ranking.
- Question: can learning kernels improve performance?