

AdaMKL: A Novel Biconvex Multiple Kernel Learning Approach

Ziming Zhang, Ze-Nian Li, Mark Drew
School of Computing Science, Simon Fraser University,
Vancouver, B.C., Canada
{zxa27, li, mark}@cs.sfu.ca

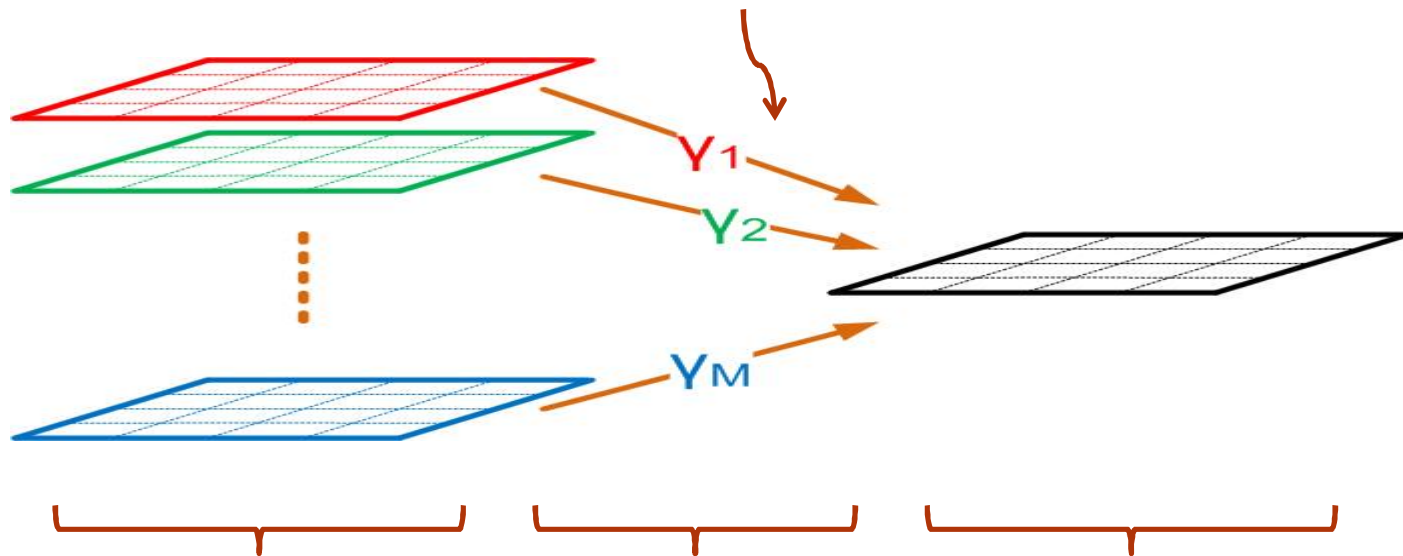
Outline

- **Background**
- **Adaptive Multiple Kernel Learning**
- **Experiments**
- **Conclusion**

Background

- Multiple Kernel Learning

- Aim to learn **kernel coefficients** and support vectors together



$$\{K_{1\dots M}\} \quad \begin{cases} \forall i, & g_i \geq 0 \\ \|g\|_p = 1, & p = 1, 2, \mathbf{L} \end{cases} \quad K_{opt} = \sum_{m=1}^M g_m K_m$$

Background

– Example: L_p-norm Multiple Kernel Learning [1]

$$\min_{\gamma, \mathbf{w}, b, \xi} \frac{1}{2} \sum_m \frac{\|\mathbf{w}_m\|_2^2}{g_m} + C \sum_i x_i$$

➔ Convex function

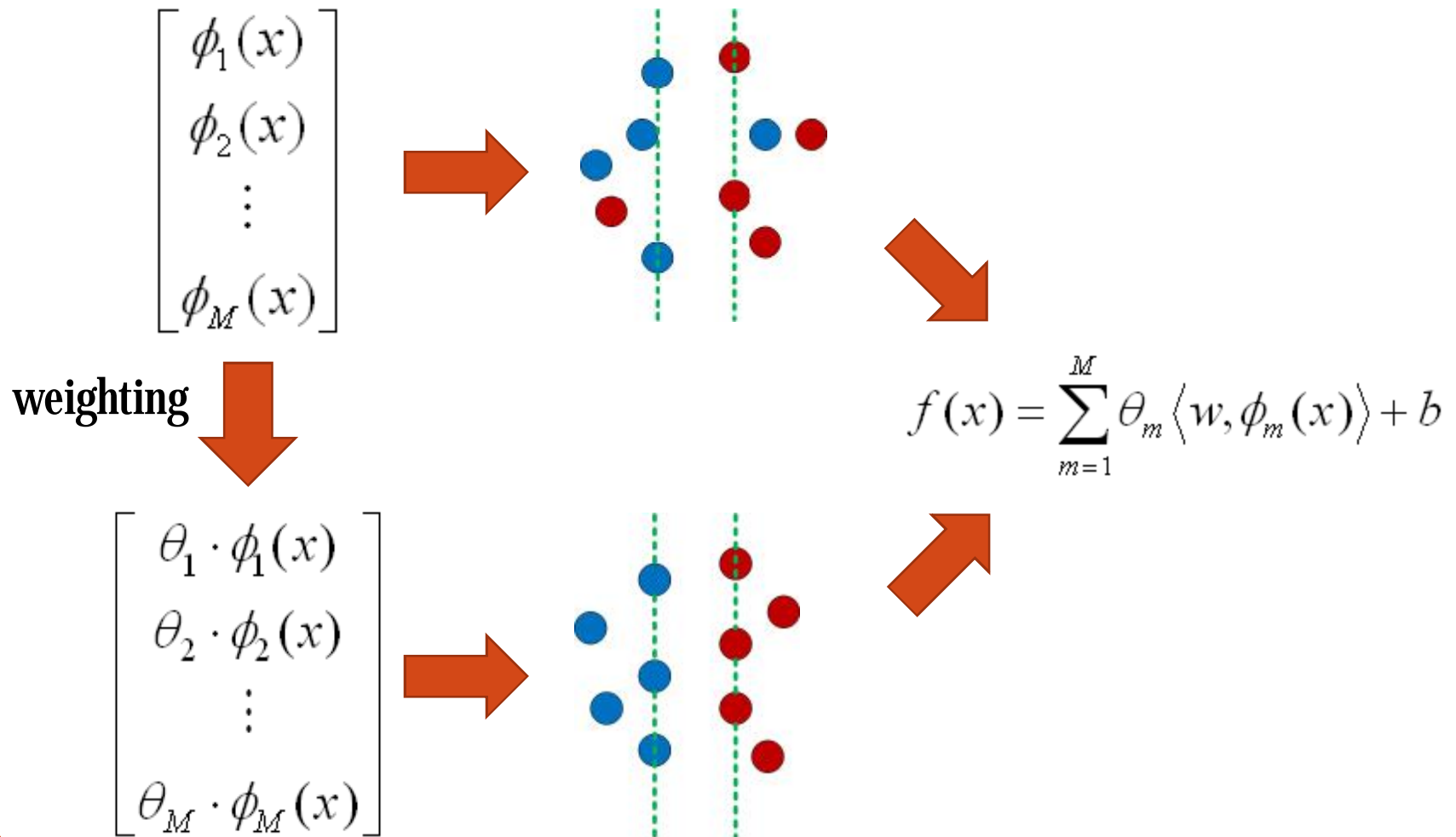
$$s.t. \quad \forall i, y_i \left[\sum_m \langle \mathbf{w}_m, \mathbf{f}_m(x_i) \rangle + b \right] \geq 1 - x_i$$
$$x_i \geq 0, C \geq 0$$

➔ Basic SVM constraints

$$\forall m, g_m \geq 0, \|g\|_p \leq 1$$

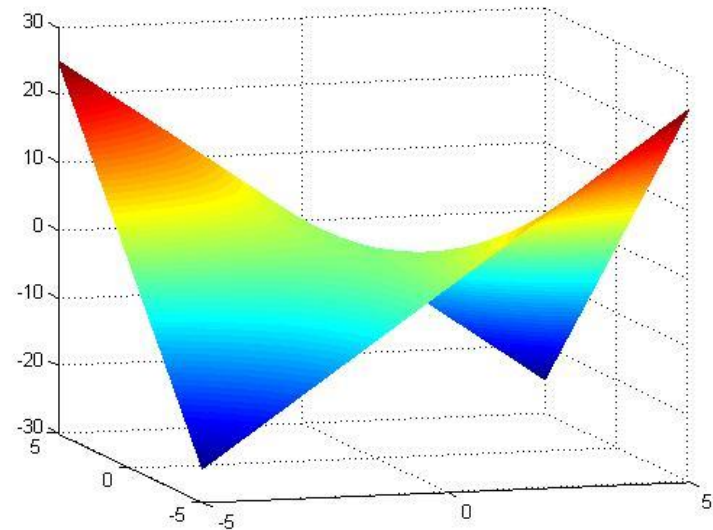
➔ Kernel coefficient constraints

Adaptive Multiple Kernel Learning



Adaptive Multiple Kernel Learning

- **Biconvex functions**
 - $f(x,y)$ is a *biconvex function* if $f_y(x)$ is convex and $f_x(y)$ is convex.
 - Example: $f(x,y)=x^2+y^2-3xy$
- **Biconvex optimization**
 - At least one function in the objective functions and constraints is biconvex.
 - Local optima



Adaptive Multiple Kernel Learning

- Adaptive Multiple Kernel Learning (AdaMKL)

- Aim to **simplify** the MKL learning process as well as keep the similar discriminative power of MKL using biconvex optimization.

- Binary classification

$$\min_{\theta, \mathbf{w}, b, \xi} \quad \frac{1}{2} N_0(\boldsymbol{\theta}) \|\mathbf{w}\|_2^2 + C \sum_i \xi_i$$

$$s.t. \quad \forall i, y_i \left[\sum_m q_m \langle \mathbf{w}_m, \mathbf{f}_m(x_i) \rangle + b \right] \geq 1 - \xi_i$$

$$\xi_i \geq 0, C \geq 0$$

$$where \quad N_0(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1^2, N_p(\boldsymbol{\theta}) = \|\boldsymbol{\theta}^2\|_{p \geq 1}$$

- Objective function: $\sum_m q_m^2 \|\mathbf{w}_m\|_2^2 \leq \|\boldsymbol{\theta}^2\|_p \|\mathbf{w}\|_2^2 \leq \|\boldsymbol{\theta}^2\|_1 \|\mathbf{w}\|_2^2 \leq \|\boldsymbol{\theta}\|_1^2 \|\mathbf{w}\|_2^2$

Adaptive Multiple Kernel Learning

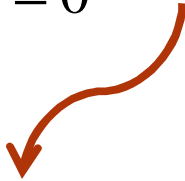
– Optimization

- Learn w by fixing θ using $N_0(\theta)$ norm
- Learn θ by fixing w using L_1 or L_2 norm of θ
- Repeat the two steps until converged

– Kernel coefficient constraints

$$\max_a \quad \sum_i a_i - \frac{1}{2} \sum_{i,j} a_i a_j y_i y_j \left[\sum_m \frac{q_m^2}{N_p(\theta)} K_m(x_i, x_j) \right] \quad (\mathbf{w}\text{-Dual})$$

$$s.t. \quad \forall i, 0 \leq a_i \leq C, \sum_i a_i y_i = 0$$


$$\mathbf{g}_m = \frac{q_m^2}{N_p(q)} \geq 0, \quad \|\mathbf{g}\|_p = 1$$

Adaptive Multiple Kernel Learning

- **Complexity**
 - Same as quadratic programming
- **Convergence**
 - If hard-margin cases ($C=+1$) can be solved at the initialization stage, then AdaMKL will converge to a local minimum.
 - If at either step our objective function converged, then AdaMKL has converged to a local minimum.

Adaptive Multiple Kernel Learning

L_p-norm MKL

$$\min_{\theta, \mathbf{w}, b, \xi} \frac{1}{2} \sum_m \frac{\|\mathbf{w}_m\|_2^2}{q_m} + C \sum_i \xi_i$$

$$s.t. \quad \forall i, y_i \left[\sum_m \langle \mathbf{w}_m, \mathbf{f}_m(x_i) \rangle + b \right] \geq 1 - \xi_i$$

$$q_m \geq 0, \|\boldsymbol{\theta}\|_p^p \leq 1$$

$$\xi_i \geq 0, C \geq 0$$

- Convex
- Kernel coefficient norm condition
- Gradient search, Semi-infinite programming (SIP), etc

AdaMKL

$$\min_{\theta, \mathbf{w}, b, \xi} \frac{1}{2} N_p(\boldsymbol{\theta}) \|\mathbf{w}\|_2^2 + C \sum_i \xi_i$$

$$s.t. \quad \forall i, y_i \left[\sum_m q_m \langle \mathbf{w}_m, \mathbf{f}_m(x_i) \rangle + b \right] \geq 1 - \xi_i$$

$$\xi_i \geq 0, C \geq 0$$

$$\max_a \sum_i a_i - \frac{1}{2} \sum_{i,j} a_i a_j y_i y_j \left[\sum_m \frac{q_m^2}{N_p(\boldsymbol{\theta})} K_m(x_i, x_j) \right]$$

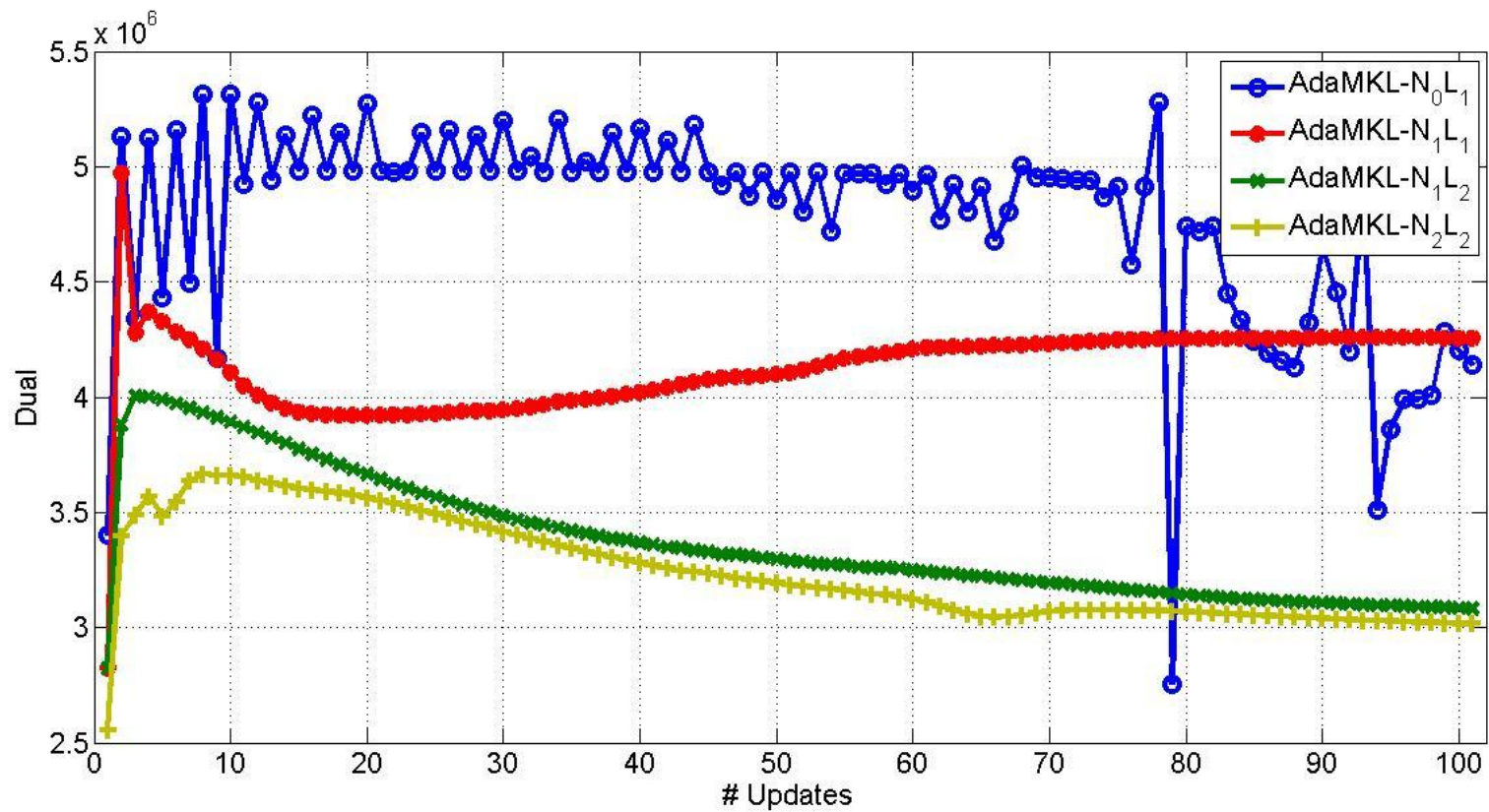
$$s.t. \quad \forall i, 0 \leq a_i \leq C, \sum_i a_i y_i = 0$$

- Biconvex
- Kernel coefficient conditions hidden in dual
- Quadratic programming

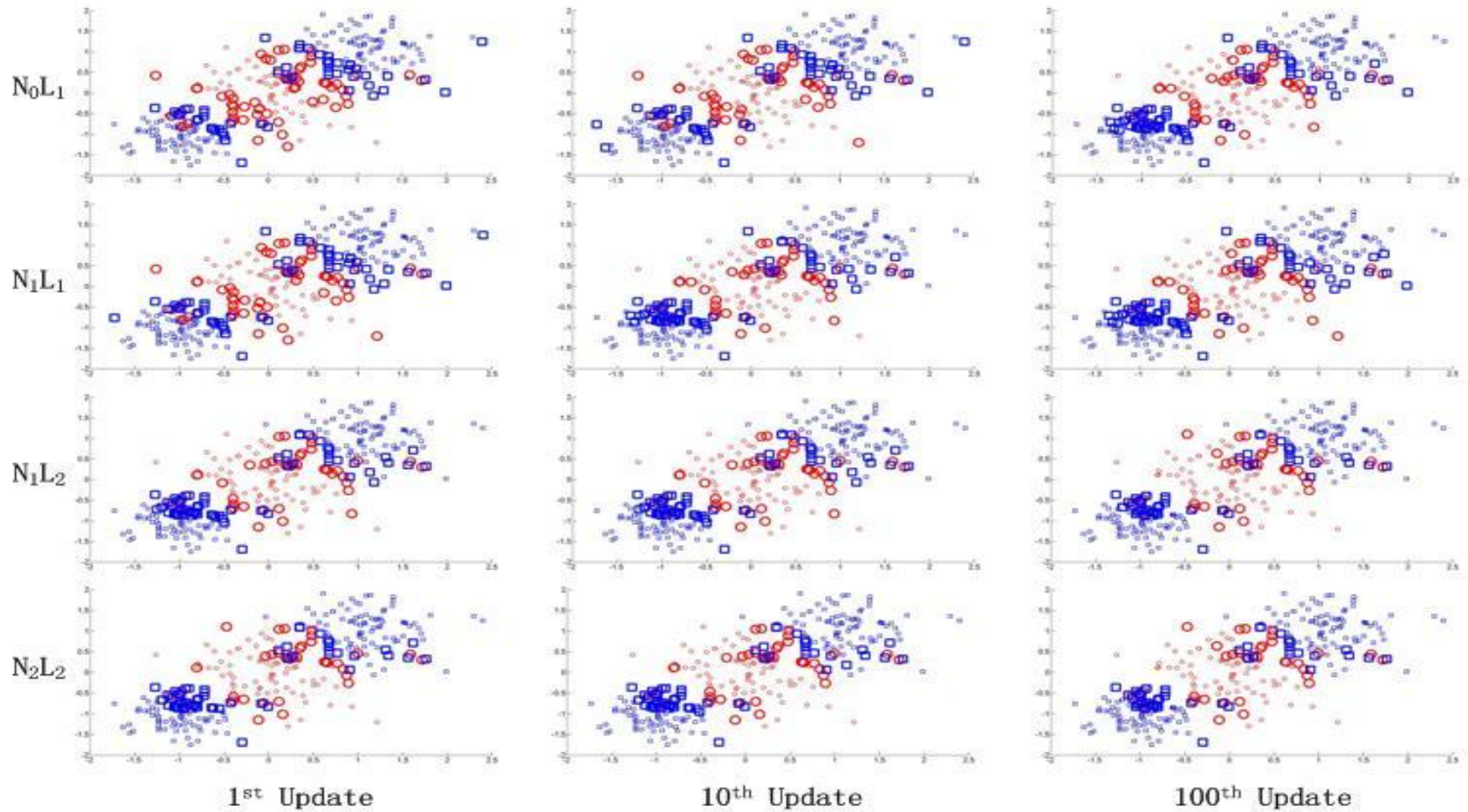
Experiments

- 4 specific AdaMKL: N_0L_1 , N_1L_1 , N_1L_2 , N_2L_2 , where “N” and “L” denote the types of norm used for learning w and θ .
- 2 experiments
 - Toy example: $C=10^5$ without tuning, 10 Gaussian kernels, randomly sampled from 2D Gaussian distributions
 - Positive samples: mean $[0\ 0]$, covariance $[0.3\ 0; 0\ 0.3]$, 100 samples
 - Negative samples: mean $[-1\ -1]$ and $[1\ 1]$, covariance $[0.1\ 0; 0\ 0.1]$ and $[0.2\ 0; 0\ 0.2]$, 100 samples, respectively.
 - 4 benchmark datasets: breast-cancer, heart, thyroid, and titanic (downloaded from <http://ida.first.fraunhofer.de/projects/bench/>)
 - Gaussian kernels + polynomial kernels
 - 100, 140, 60, 40 kernels for corresponding datasets, respectively

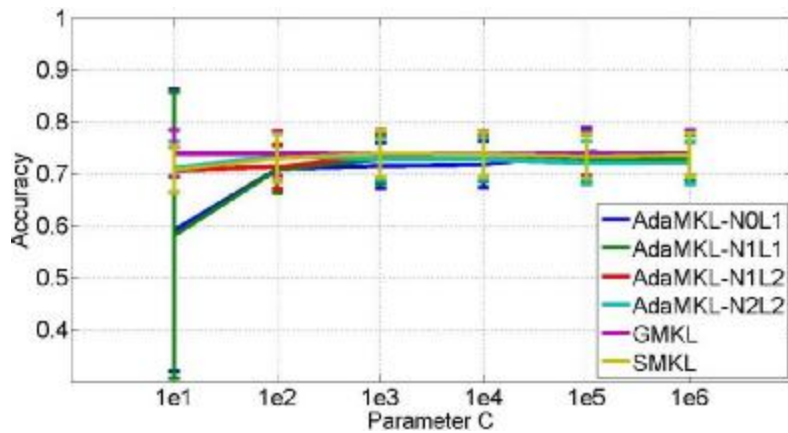
Experiments - Toy example



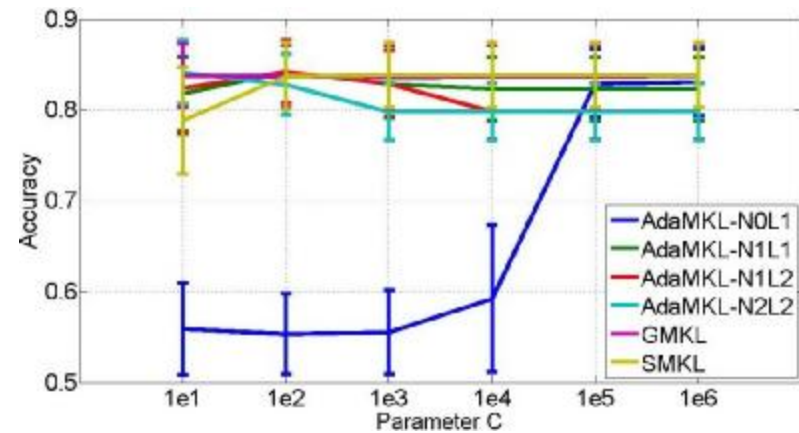
Experiments - Toy example



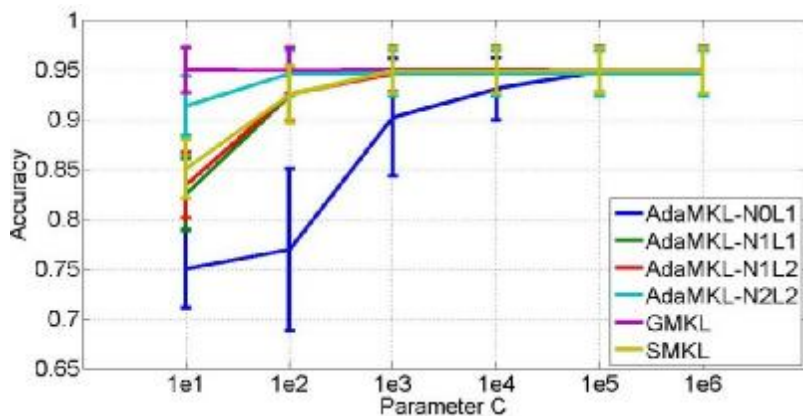
Experiments - Benchmark datasets



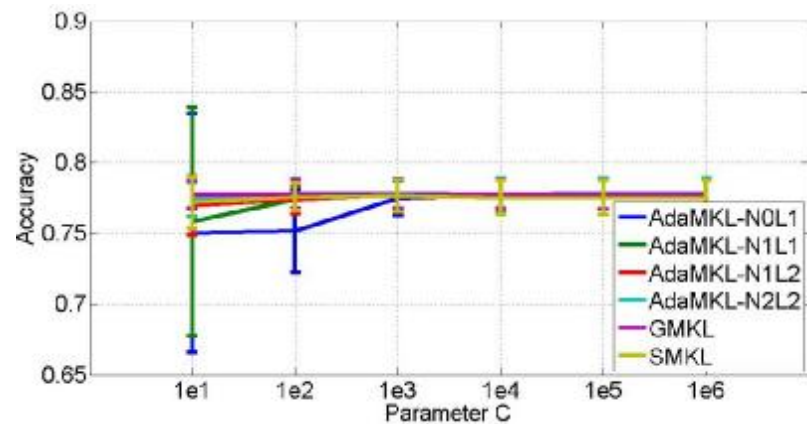
(a) Breast-Cancer: [69.64 ~ 75.23]



(b) Heart: [79.71 ~ 84.05]

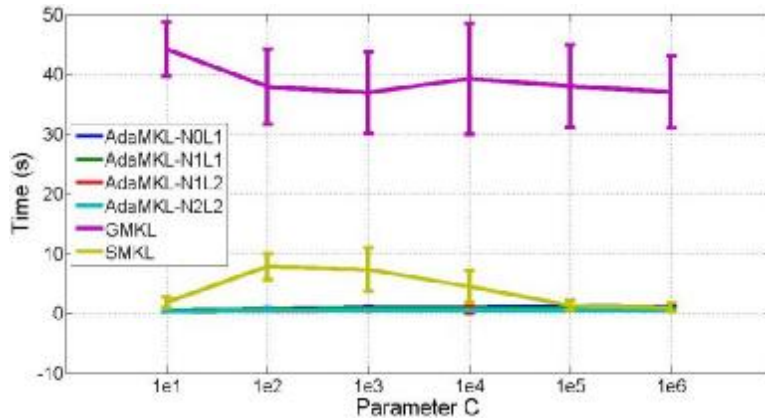


(c) Thyroid: [95.20 ~ 95.80]

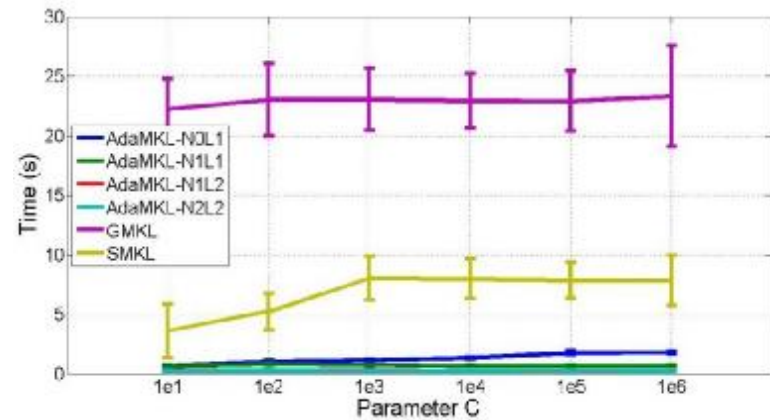


(d) Titanic: [76.02 ~ 77.58]

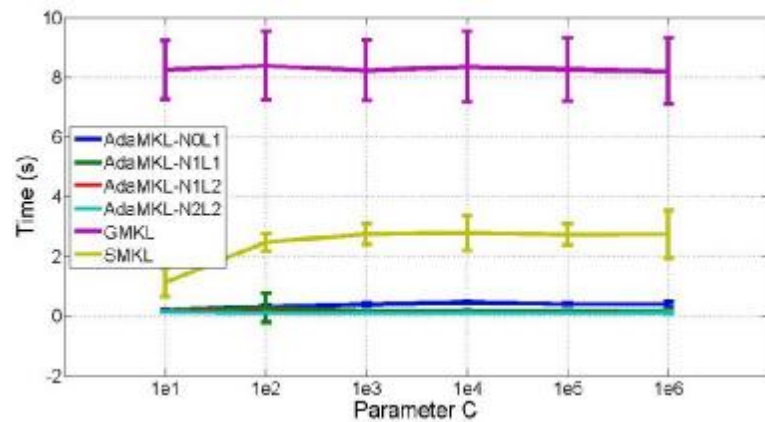
Experiments - Benchmark datasets



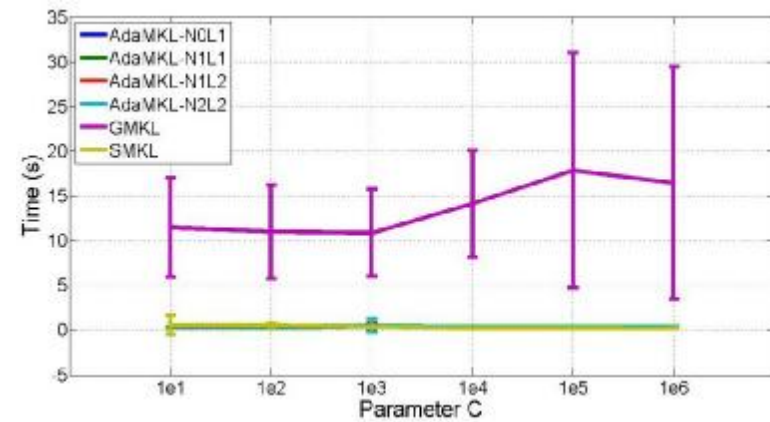
(a) Breast-Cancer



(b) Heart



(c) Thyroid



(d) Titanic

Conclusion

- **Biconvex optimization for MKL**
 - **Arbitrary L_p norm of kernel coefficient constraint, which is actually hidden in the dual without consideration explicitly.**
 - **Easy to optimize, fast to converge, lower computational time but similar performance as traditional convex optimization based MKL**

Thank you !!!