Optimization for Machine Learning

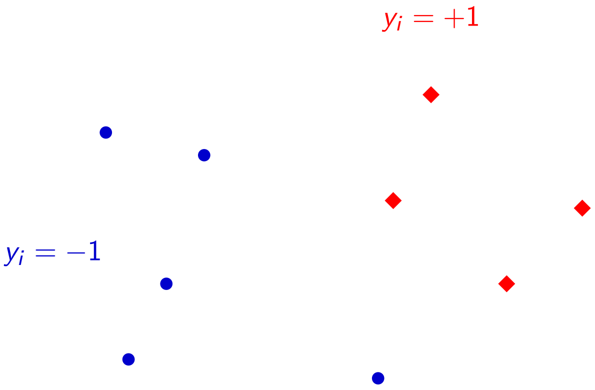
Lecture 4: SMO-MKL

S.V. N. (vishy) Vishwanathan

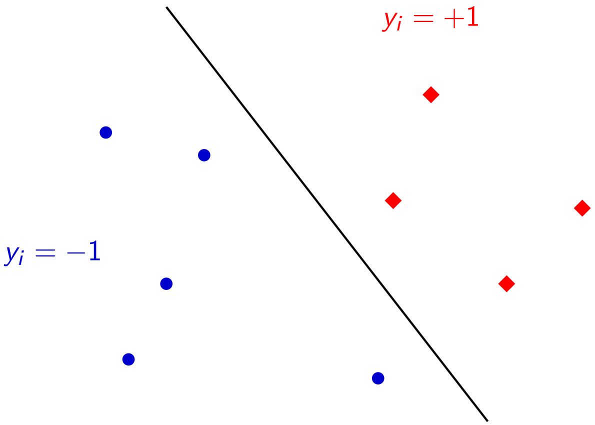
Purdue University
vishy@purdue.edu

July 11, 2012

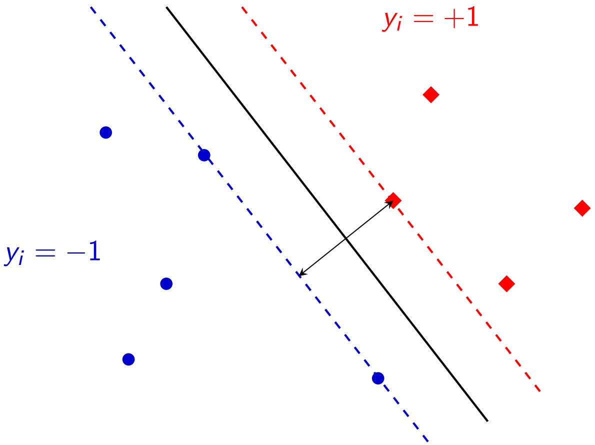
Binary Classification



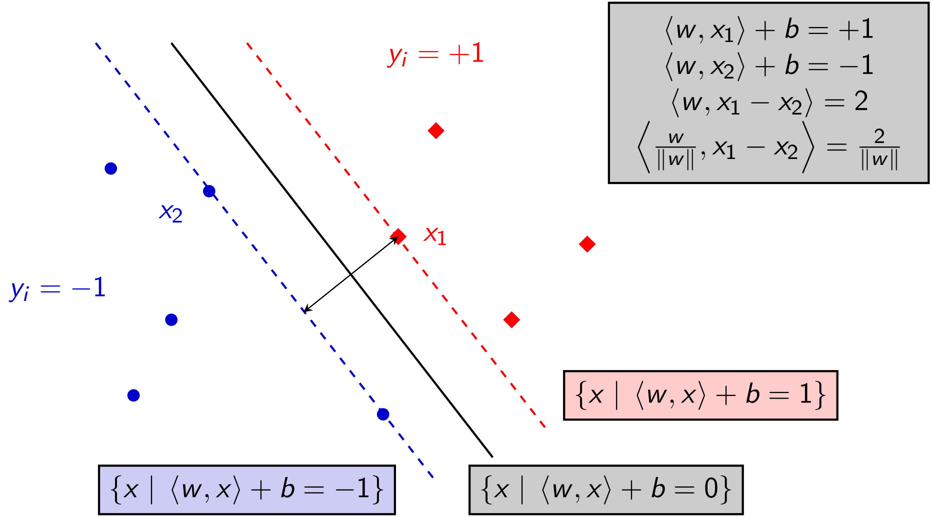
Binary Classification



Binary Classification



Binary Classification

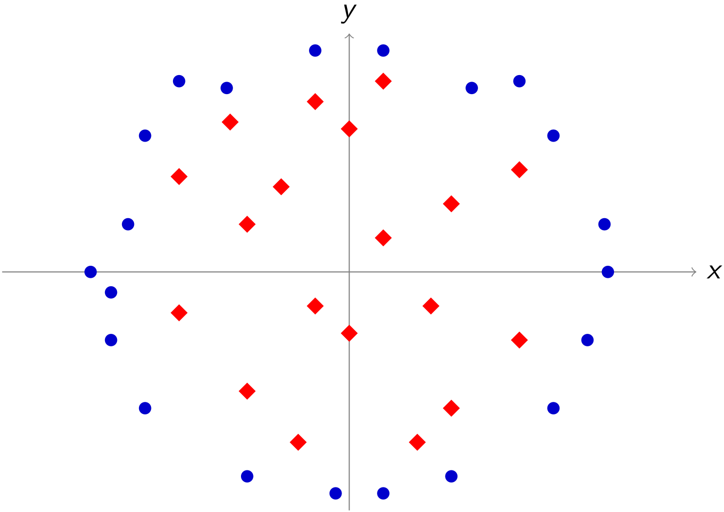


Linear Support Vector Machines

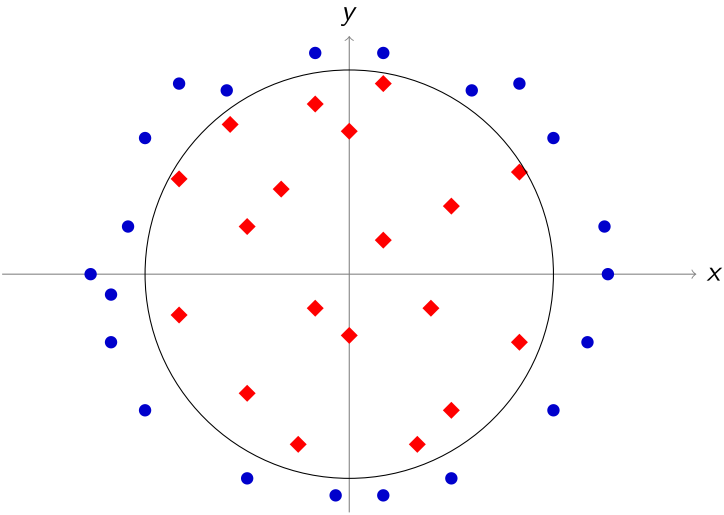
Optimization Problem

$$
\begin{aligned}
\min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \\
\text{s.t.} \quad & y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i \text{ for all } i \\
& \xi_i \geq 0
\end{aligned}
$$

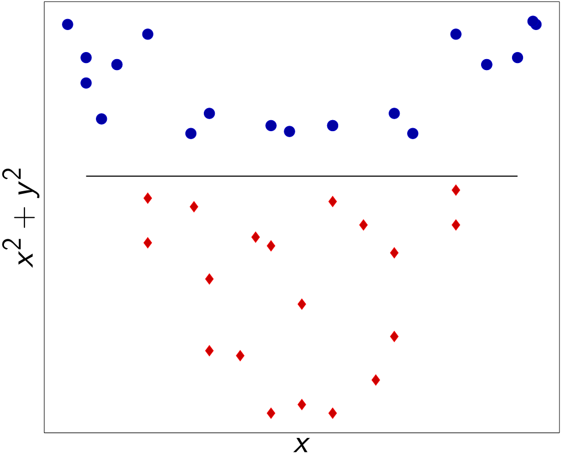
The Kernel Trick



The Kernel Trick



The Kernel Trick



Kernel Trick

Optimization Problem

$$
\begin{aligned}
\min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i \\
\text{s.t.} \quad & y_i(\langle w, \phi(x_i)\rangle + b) \geq 1 - \xi_i \text{ for all } i \\
& \xi_i \geq 0
\end{aligned}
$$

Kernel Trick

Optimization Problem

$$\max_{\alpha} \quad -\frac{1}{2}\alpha^\top H\alpha + \mathbf{1}^\top \alpha$$

$$\text{s.t.} \quad 0 \leq \alpha_i \leq C$$

$$\sum_i \alpha_i y_i = 0$$

$$H_{ij} = y_i y_j \langle \phi(x_i), \phi(x_j) \rangle$$

Kernel Trick

Optimization Problem

$$\max_{\alpha} \quad -\frac{1}{2}\alpha^\top H\alpha + \mathbf{1}^\top \alpha$$
$$\text{s.t.} \quad 0 \leq \alpha_i \leq C$$
$$\sum_i \alpha_i y_i = 0$$

$$H_{ij} = y_i y_j k(x_i, x_j)$$

Key Question

Which kernel should I use?

The Multiple Kernel Learning Answer

- Cook up as many (base) kernels as you can
- Compute a data dependent kernel function as a linear combination of base kernels

$$k(x, x') = \sum_k d_k k_k(x, x') \quad \text{s.t. } d_k \geq 0$$

Key Question

Which kernel should I use?

The Multiple Kernel Learning Answer

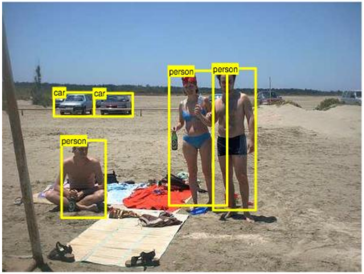
- Cook up as many (base) kernels as you can
- Compute a data dependent kernel function as a linear combination of base kernels

$$k(x, x') = \sum_k d_k k_k(x, x') \quad \text{s.t. } d_k \geq 0$$

Object Detection

Localize a specified object of interest if it exists in a given image

Some Examples of MKL Detection



Summary of Our Results

- Sonar Dataset with 800 kernels

p	Training Time (s)		# Kernels Selected	
	SMO-MKL	Shogun	SMO-MKL	Shogun
1.1	4.71	47.43	91.20	258.00
1.33	3.21	19.94	248.20	374.20
2.0	3.39	34.67	661.20	664.80

- Web dataset: $\approx$ 50,000 points and 50 kernels $\approx$ 30 minutes
- Sonar with a hundred thousand kernels
  - Precomputed: $\approx$ 8 minutes
  - Kernels computed on-the-fly: $\approx$ 30 minutes

Setting up the Optimization Problem -I

The Setup

- We are given K kernel functions $k_1, \ldots, k_n$ with corresponding feature maps $\phi_1(\cdot), \ldots, \phi_n(\cdot)$
- We are interested in deriving the feature map

$$\phi(x) = \begin{bmatrix} \sqrt{d_1}\phi_1(x) \\ \vdots \\ \sqrt{d_n}\phi_n(x) \end{bmatrix}$$

Setting up the Optimization Problem -I

The Setup

- We are given K kernel functions $k_1, \ldots, k_n$ with corresponding feature maps $\phi_1(\cdot), \ldots, \phi_n(\cdot)$
- We are interested in deriving the feature map

$$\phi(x) = \begin{bmatrix} \sqrt{d_1}\phi_1(x) \\ \vdots \\ \sqrt{d_n}\phi_n(x) \end{bmatrix} \implies w = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}$$

Setting up the Optimization Problem

Optimization Problem

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & y_i(\langle w, \phi(x_i)\rangle + b) \geq 1 - \xi_i \text{ for all } i \\ & \xi_i \geq 0 \end{aligned}$$

Setting up the Optimization Problem

Optimization Problem

$$\min_{w,b,\xi,d} \quad \frac{1}{2}\sum_k \|w_k\|^2 + C\sum_{i=1}^m \xi_i$$

$$\text{s.t.} \quad y_i\left(\sum_k \sqrt{d_k}\,\langle w_k, \phi_k(x_i)\rangle + b\right) \geq 1 - \xi_i \text{ for all } i$$

$$\xi_i \geq 0$$

$$d_k \geq 0 \text{ for all } k$$

Setting up the Optimization Problem

Optimization Problem

$$
\begin{aligned}
\min_{w,b,\xi,d} \quad & \frac{1}{2}\sum_k \|w_k\|^2 + C\sum_{i=1}^m \xi_i + \frac{\rho}{2}\left(\sum_k d_k^p\right)^{\frac{2}{p}} \\
\text{s.t.} \quad & y_i\left(\sum_k \sqrt{d_k}\,\langle w_k, \phi_k(x_i)\rangle + b\right) \geq 1 - \xi_i \text{ for all } i \\
& \xi_i \geq 0 \\
& d_k \geq 0 \text{ for all } k
\end{aligned}
$$

Setting up the Optimization Problem

Optimization Problem

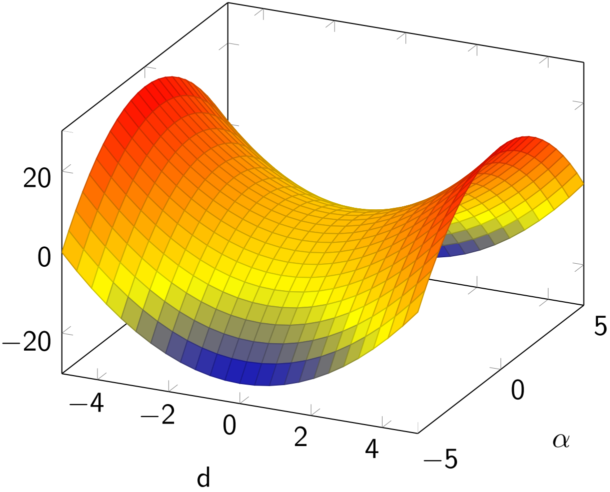
$$
\begin{aligned}
\min_{w,b,\xi,d} \quad & \frac{1}{2}\sum_{k}\frac{\left\|w_k\right\|^2}{d_k} + C\sum_{i=1}^{m}\xi_i + \frac{\rho}{2}\left(\sum_{k} d_k^p\right)^{\frac{2}{p}} \\
\text{s.t.} \quad & y_i\left(\sum_{k}\langle w_k, \phi_k(x_i)\rangle + b\right) \geq 1 - \xi_i \text{ for all } i \\
& \xi_i \geq 0 \\
& d_k \geq 0 \text{ for all } k
\end{aligned}
$$

Setting up the Optimization Problem

Optimization Problem

$$
\begin{aligned}
\min_{d} \max_{\alpha} \quad & -\frac{1}{2} \sum_{k} d_k \alpha^\top H_k \alpha + \mathbf{1}^\top \alpha + \frac{\rho}{2} \left(\sum_{k} d_k^p \right)^{\frac{2}{p}} \\
\text{s.t.} \quad & 0 \leq \alpha_i \leq C \\
& \sum_{i} \alpha_i y_i = 0 \\
& d_k \geq 0
\end{aligned}
$$

Saddle Point Problem



Solving the Saddle Point

Saddle Point Problem

$$\min_{d} \max_{\alpha} \quad -\frac{1}{2}\sum_{k} d_k \alpha^\top H_k \alpha + \mathbf{1}^\top \alpha + \frac{\rho}{2}\left(\sum_{k} d_k^p\right)^{\frac{2}{p}}$$

$$\text{s.t.} \quad 0 \leq \alpha_i \leq C$$

$$\sum_{i} \alpha_i y_i = 0$$

$$d_k \geq 0$$

The Key Insight

Eliminate d

$$\begin{aligned} D(\alpha) := \max_{\alpha} \quad & -\frac{1}{8\rho}\left(\sum_k \left(\alpha^\top H_k \alpha\right)^q\right)^{\frac{2}{q}} + \mathbf{1}^\top \alpha \\ \text{s.t.} \quad & 0 \le \alpha_i \le C \\ & \sum_i \alpha_i y_i = 0 \\ & \frac{1}{p} + \frac{1}{q} = 1 \end{aligned}$$

Not a QP but very close to one!

SMO-MKL: High Level Overview

$$D(\alpha) := \max_{\alpha} \; -\frac{1}{8\rho}\left(\sum_k \left(\alpha^\top H_k \alpha\right)^q\right)^{\frac{2}{q}} + \mathbf{1}^\top \alpha$$
$$\text{s.t.} \quad 0 \le \alpha_i \le C$$
$$\sum_i \alpha_i y_i = 0$$

Algorithm

- Choose two variables α_i and α_j to optimize
- Solve the one dimensional reduced optimization problem
- Repeat until convergence

SMO-MKL: High Level Overview

Selecting the Working Set

- Compute directional derivative and directional Hessian
- Greedily select the variables

Solving the Reduced Problem

- Analytic solution for $p = q = 2$ (one dimensional quartic)
- For other values of p use Newton Raphson

SMO-MKL: High Level Overview

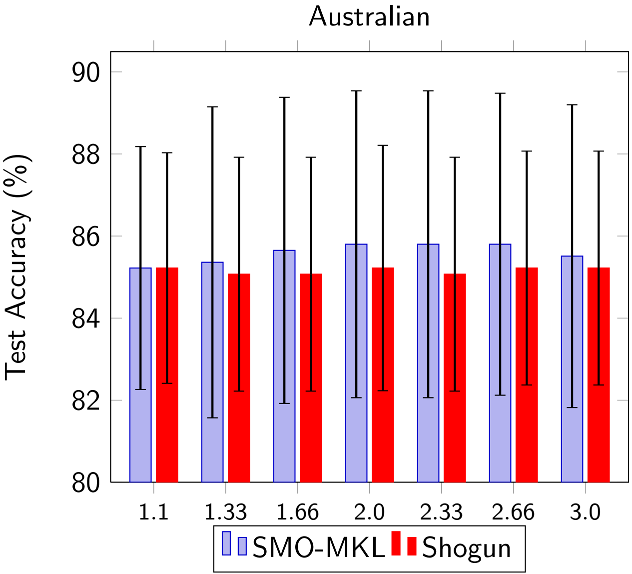
Selecting the Working Set

- Compute directional derivative and directional Hessian
- Greedily select the variables

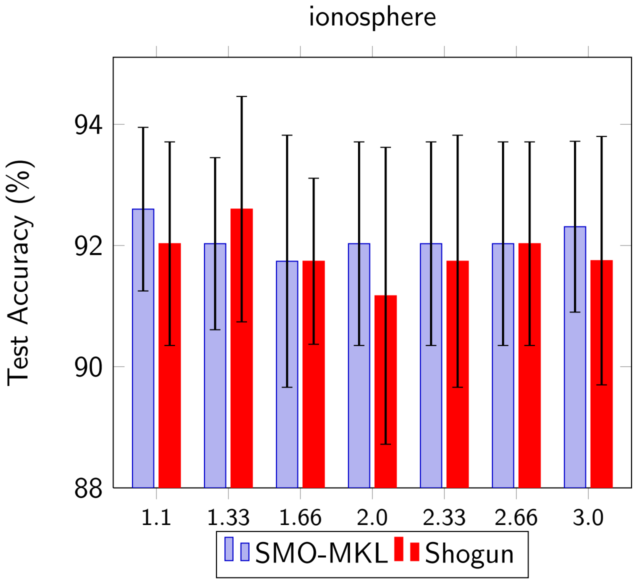
Solving the Reduced Problem

- Analytic solution for $p = q = 2$ (one dimensional quartic)
- For other values of p use Newton Raphson

Generalization Performance

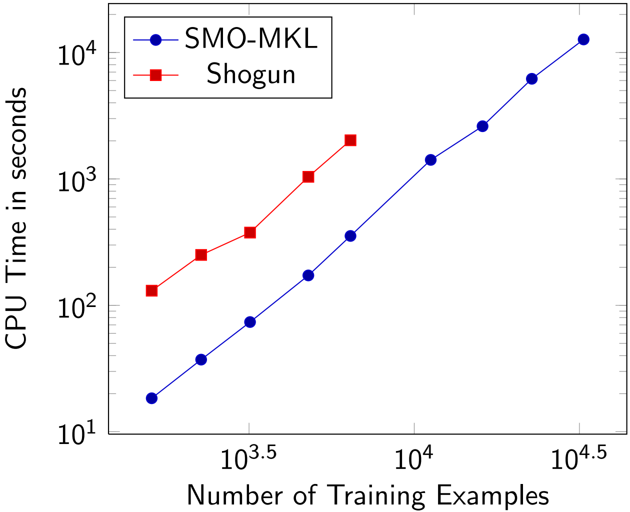


Generalization Performance



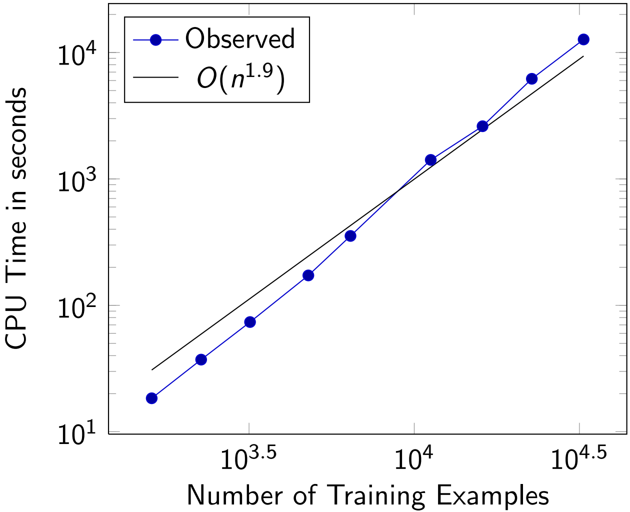
Scaling with Training Set Size

Adult: 123 dimensions, 50 RBF kernels, $p = 1.33$, $C = 1$



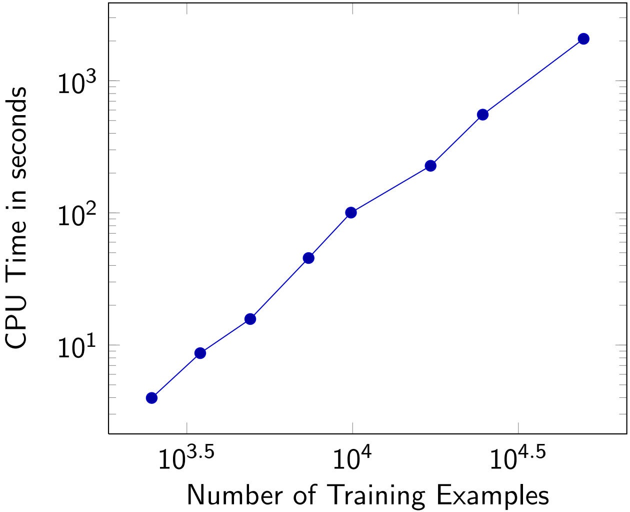
Scaling with Training Set Size

Adult: 123 dimensions, 50 RBF kernels, $p = 1.33$, $C = 1$



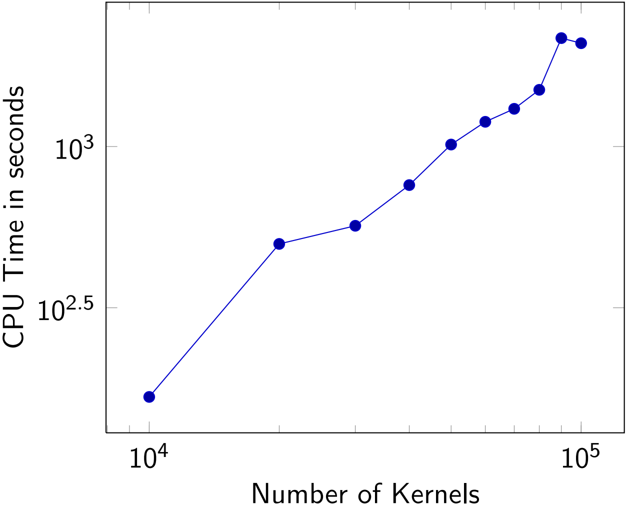
On Another Dataset

Web: 300 dimensions, 50 RBF kernels, $p = 1.33$, $C = 1$



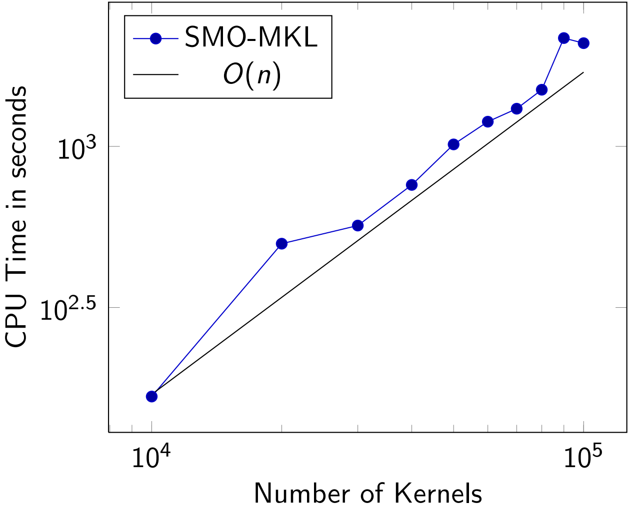
Scaling with Number of Kernels

Sonar: 208 examples, 59 dimensions, $p = 1.33$, $C = 1$



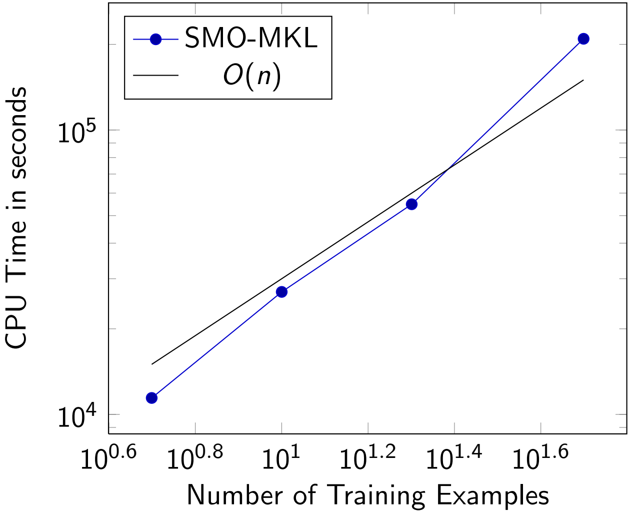
Scaling with Number of Kernels

Sonar: 208 examples, 59 dimensions, $p = 1.33$, $C = 1$



On Another Dataset

Real-sim: 72,309 examples, 20,958 dimensions, $p = 1.33$, $C = 1$



References

- S. V. N. Vishwanathan, Zhaonan Sun, Theera-Ampornpunt, and Manik Varma. *Multiple Kernel Learning and the SMO Algorithm*. NIPS 2010. Pages 2361–2369.
- Code available for download from `http://research.microsoft.com/en-us/um/people/manik/code/SMO-MKL/download.html`