KDD 2012 BEIJING
THE 18TH ACM SIGKDD CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING
Beijing, China August 12-16, 2012

SPG-GMKL: Generalized Multiple Kernel Learning with a Million Kernels

Ashesh Jain
IIT Delhi
asheshjain399@gmail.com

S. V. N. Vishwanathan
Purdue University
vishy@stat.purdue.edu

Manik Varma
Microsoft Research India
manik@microsoft.com

1. Kernel Learning

•Jointly learning both SVM and kernel parameters from training data
• Kernel parameterizations
  • Linear : $K = \sum_i d_i K_i$
  • Non-linear : $K = \prod_i K_i = \prod_i e^{-d_i D_i}$
• Regularizers
  • Sparse l_1
  • Sparse and non-sparse $l_{p>1}$

2. GMKL Formulation for binary classification

$P = \text{Min}_{\mathbf{w},b,\mathbf{d},\xi}\ \frac{1}{2}\mathbf{w}^t\mathbf{w} + C\sum_i \xi_i + r(\mathbf{d})$
s. t. $y_i(\mathbf{w}^t\boldsymbol{\phi}_{\mathbf{d}}(\mathbf{x}_i) + b) \geq 1 - \xi_i$
$\xi_i \geq 0$ **&** $\mathbf{d} \in \boldsymbol{D}$

Intermediate Dual:

$D = \text{Min}_{\mathbf{d}}\text{Max}_{\alpha}\ \mathbf{1}^t\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^t\mathbf{Y}\mathbf{K}_{\mathbf{d}}\mathbf{Y}\boldsymbol{\alpha} + r(\mathbf{d})$
s. t. $\mathbf{1}^t\mathbf{Y}\boldsymbol{\alpha} = 0$
$\mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{C}$ **&** $\mathbf{d} \in \boldsymbol{D}$

• Can be extended to other loss functions and regularizers

3. Wrapper method for optimizing GMKL

Outer loop: $\text{Min}_{\mathbf{d}}\ W(\mathbf{d})$ s. t. $\mathbf{d} \in \boldsymbol{D}$

Inner loop: $W(\mathbf{d}) = \text{Max}_{\boldsymbol{\alpha}}\ \mathbf{1}^t\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^t\mathbf{Y}\mathbf{K}_{\mathbf{d}}\mathbf{Y}\boldsymbol{\alpha} + r(\mathbf{d})$
s. t. $\mathbf{1}^t\mathbf{Y}\boldsymbol{\alpha} = 0$ **&** $\mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{C}$

• $\nabla_{\mathbf{d}} W = -\frac{1}{2}\boldsymbol{\alpha}^{*t}\mathbf{Y}\,\nabla_{\mathbf{d}}\mathbf{K}\mathbf{Y}\boldsymbol{\alpha}^* + \nabla_{\mathbf{d}} r$
• $\boldsymbol{\alpha}^*$ can be obtained using a standard SVM solver
• Optimized using Projected Gradient Descent (PGD)

4. PGD limitations

• Requires many function and gradient evaluations as
  • No step size information is available
  • Armijo rule might reject many step size proposals
  • Noisy gradient can lead to many tiny steps
• Solving SVMs to high precision to obtain accurate function and gradient values is very expensive
• Repeated projection onto the feasible set can be expensive

5. SPG Solution

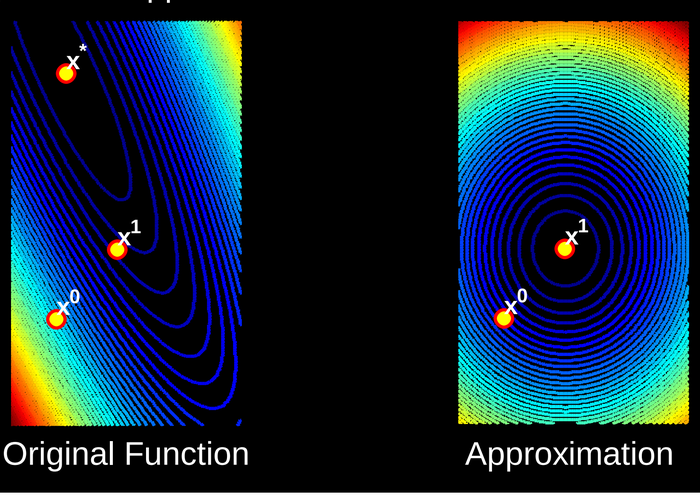
SPG is obtained by adding four components to PGD
1. Spectral Step Length
$$\lambda_{SPG} = \frac{\langle \mathbf{d}^n - \mathbf{d}^{n-1},\ \mathbf{d}^n - \mathbf{d}^{n-1}\rangle}{\langle \mathbf{d}^n - \mathbf{d}^{n-1},\ \nabla W(\mathbf{d}^n) - \nabla W(\mathbf{d}^{n-1})\rangle}$$
2. Non-monotone line search
$$W(\mathbf{d}^n - s\nabla W(\mathbf{d}^n)) \leq \max_{0\leq j\leq M} W(\mathbf{d}^{n-j}) - \gamma s|\nabla W(\mathbf{d}^n)|_2^2$$
3. Inner SVM precision tuning
   • Quicker steps when away from minima
4. Minimizing the number of projections

6. SPG Algorithm

$n \leftarrow 0$; Initialize $\mathbf{d}^0$ randomly
repeat
  $\boldsymbol{\alpha}^* \leftarrow \text{SolveSVM}_{\epsilon}(\mathbf{K}(\mathbf{d}^n))$
  $\lambda \leftarrow \text{SpectralStepLength}$
  $\mathbf{p}^n \leftarrow \mathbf{d}^n - \mathbf{P}(\mathbf{d}^n - \boldsymbol{\lambda}\nabla W(\mathbf{d}^n, \boldsymbol{\alpha}^*))$
  $s^n \leftarrow \text{Non} - \text{Monotone}$
  $\epsilon \leftarrow \text{TuneSVMPrecision}$
  $\mathbf{d}^{n+1} \leftarrow \mathbf{d}^n - s^n\mathbf{p}^n$
until converged

7. Spectral Step Length

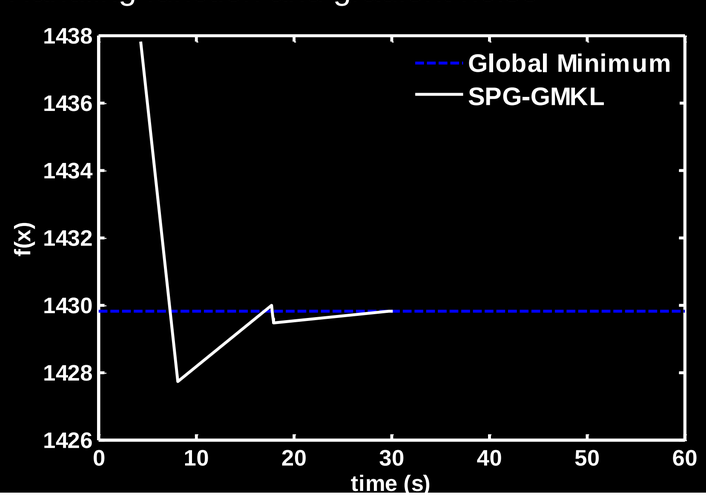
• Quadratic approximation : $\frac{1}{2}\lambda\mathbf{x}^t\mathbf{x} + \mathbf{c}^t\mathbf{x} + d$



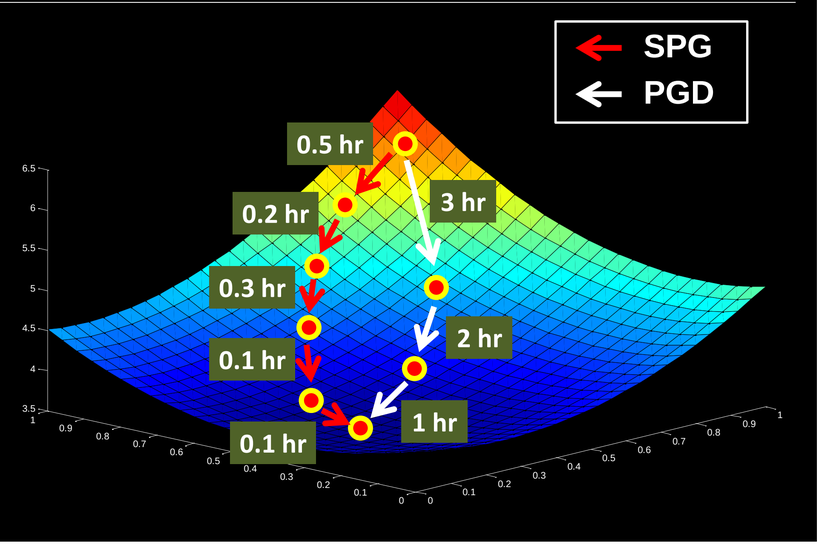
Original Function Approximation

8. Non-Monotone Rule

• Handling function and gradient noise



9. SVM Precision Tuning



10. Results on Large Scale Data Sets

• Sum of kernels subject to $l_{p\geq 1}$ regularization

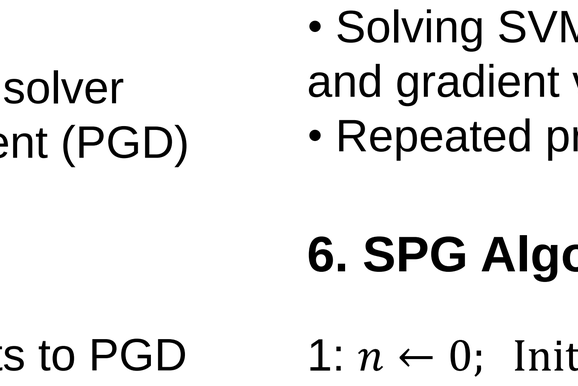
Data Sets	# Train	# Kernels	l_1		$l_{1.33}$	
			PGD (hrs)	SPG (hrs)	PGD (hrs)	SPG (hrs)
Adult - 9	32,561	50	35.84	4.55	31.77	4.42
Cod - RNA	59,535	50	–	25.17	66.48	19.10
KDDCup04	50,000	50	–	40.10	–	42.20
Sonar	208	1 Million			–	105.62
Covertype	581,012	5			–	64.46

• Product of kernels subject to $l_{p\geq 1}$ regularization

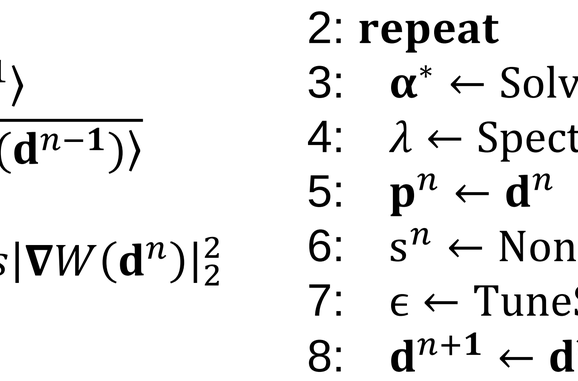
Data Sets	# Train	# Kernels	l_1		$l_{1.33}$	
			PGD (hrs)	SPG (hrs)	PGD (hrs)	SPG (hrs)
Letter	20,000	16	18.66	0.67	18.69	0.66
Poker	25,010	10	5.57	0.49	2.29	0.96
Adult - 8	22,696	42	–	1.73	–	3.42
Web - 7	24,692	43	–	0.88	–	1.33
RCV1	20,242	50	–	18.17	–	15.93
Cod - RNA	59,535	8	–	3.45	–	8.99

11. SPG Scaling Properties

• Scaling with the number of training points



• Scaling with the number of kernels



12. Results on Small Scale Data Sets

• Sum of kernels subject to l_1 regularization

Data Sets	SimpleMKL (s)	Shogun (s)	PGD (s)	SPG (s)
Wpbc	400 ± 356.4	15 ± 7.7	38 ± 17.6	6 ± 4.2
Breast - Cancer	676 ± 356.4	12 ± 1.2	57 ± 85.1	5 ± 0.6
Australian	383 ± 33.5	1094 ± 621.6	29 ± 7.1	10 ± 0.8
Ionosphere	1247 ± 680.0	107 ± 18.8	1392 ± 824.2	39 ± 6.8
Sonar	1468 ± 1252.7	935 ± 65.0	–	273 ± 64.0

13. SPG Contributions

• SPG can handle any regularization and kernel parameterization
• SPG is more robust to noisy function and gradient values
• SPG requires fewer function and gradient evaluations

Paper ID: rt540