Comparing Sparse and Non-Sparse Multiple Kernel Learning

M. Kloft[1], U. Brefeld[2], S. Sonnenburg[3], and A. Zien[4]

[1] University of California, Berkeley, USA
[2] Yahoo Research, Barcelona, Spain
[3] Technical University of Berlin, Germany
[4] LIFE Biosystems GmbH, Heidelberg, Germany

Contributions

- Generalization of [1] to arbitrary convex loss functions and arbitrary norms.
- *Simple* optimization procedure based on an analytical update of the kernel weights.
- Toy experiment gives insight in the trade-off between sparsity and accuracy in sparse and non-sparse scenarios.
- New *large-scale* runtime experiments show efficiency of interleaved optimization approaches.

Generalized Multiple Kernel Learning

- Given data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n) \subset \mathcal{X} \times \mathcal{Y}$.
- RKHS feature mappings $\psi_m : \mathcal{X} \to \mathcal{H}_m$ and Hilbertian norms $||\cdot||_{\mathcal{H}_m}$.
- An arbitrary convex loss function V
- Consider generalized MKL problem:

$$\inf_{\boldsymbol{w},b,\boldsymbol{\theta}:\boldsymbol{\theta}\geq 0} \quad C\sum_{i=1}^{n} V\left(\sum_{m=1}^{M}\langle \boldsymbol{w}_m, \psi_m(\boldsymbol{x}_i)\rangle_{\mathcal{H}_m} + b,\ y_i\right) + \frac{1}{2}\sum_{m=1}^{M}\frac{\|\boldsymbol{w}_m\|_{\mathcal{H}_m}^2}{\theta_m}$$
$$\text{s.t.} \quad \|\boldsymbol{\theta}\|^2 \leq 1, \tag{1}$$

where $||\cdot||$ is an arbitrary norm.

- Solving for optimal w and b and resubstitute into $\mathcal{L}$.
- Incorporate t and θ into Fenchel-Legendre conjugates.
- Obtain *generalized dual*:

$$\sup_{\boldsymbol{\alpha}:\mathbf{1}^\top\boldsymbol{\alpha}=0} \quad -C\sum_{i=1}^{n} V^*\left(-\frac{\alpha_i}{C},\ y_i\right) - \frac{1}{2}\left\|(\boldsymbol{\alpha}K_m\boldsymbol{\alpha})_{m=1}^{M}\right\|_*,$$

with Fenchel-Legendre conjugates V^* and $||\cdot||_*$.

"Direct" Optimization

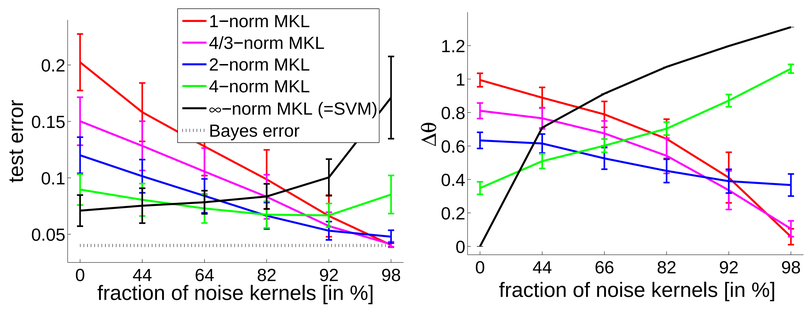
- Consider case where $||\cdot||$ is an $\ell_{p>1}$-norm.
- Solving for $\frac{\partial \mathcal{L}}{\partial \theta_m} = 0$ gives

$$\theta_m = \frac{\left(||w_m||^2\right)^{\frac{1}{p+1}}}{\left(\sum_{m=1}^{M}\left(||w_m||^2\right)^{\frac{p}{p+1}}\right)^{1/p}}. \tag{2}$$

- Dual prohibits efficient optimization.
- Instead, alternate optimization of (1):
  - *Repeat*
  - $(\boldsymbol{w}, b)$-step: optimization wrt $(\boldsymbol{w}, b)$ (regular SVM training).
  - $\boldsymbol{\theta}$-step: analytic update according to (2)
  - *until convergence*

Toy Experiment

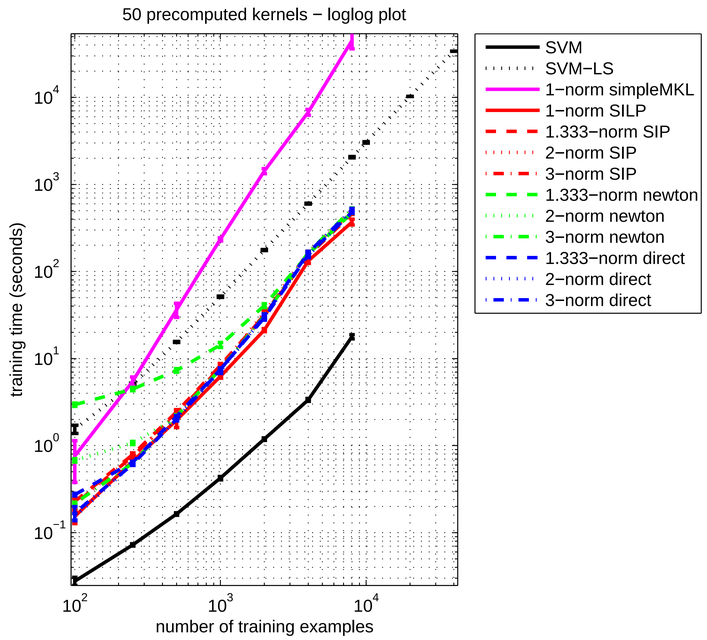
- Binary labeled data $(\boldsymbol{x}_i, y_i)$ from two 50-dim isotropic Gaussians.
- Gaussians' center: $\mu_1 = \rho\boldsymbol{\theta}$, $\mu_2 = -\rho\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is a binary vector.
- Each feature normalized (std=1) and processed by linear kernel.
- Fraction of noise kernels is defined as $\nu(\boldsymbol{\theta}) = \frac{1}{d}||\boldsymbol{\theta}||_1$.
- 50 training, 5,000 validation, 10,000 hold out set. $p = 1, \frac{4}{3}, 2, 4, \infty$.



- 2-norm MKL best prediction model in our experiments.
- Sparse MKL inferior when the noise level is between 0-92%.

Execution Times

- 50 RBF kernels; sample size varied.
- SVM, SimpleMKL [2], ℓ^p-norm MKL with $p \in \{1, 1.333, 2, 3, \infty\}$.
- Averages and standard errors over 5 repetitions.



- Execution time of ℓ^p-norm MKL depends on parameter p.
- All interleaved optimization methods convergence quickly!
- SimpleMKL outperformed by two orders of magnitude.

References

[1] M. Kloft, U. Brefeld, S. Sonnenburg, P. Laskov, K.-R. Müller, and A. Zien. Efficient and Accurate ℓ_p-norm MKL, NIPS 22, (to appear) 2010.

[2] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. JMLR (9), 2491-2521, 2008.