Details on Optimizing P1

1 Optimization of P1

For notation convenience, we switch the symbols v_{m+1} and w_{m+1} in the original **P1**. Then **P1** in the submitted paper becomes

$$\begin{aligned}
\textbf{P1}\quad & \min_{\boldsymbol{w},\boldsymbol{v},\boldsymbol{\xi},\boldsymbol{d}} \frac{1}{2}\sum_{l=1}^{m+1} d_l\|v_l\|^2 + \frac{1}{2}\sum_{l=1}^{m} C_l d_l\|w_l\|^2 + \frac{1}{2}\sum_{l=1}^{m}\sum_{i=1}^{n} d_l(\xi_i^l)^2 + C\sum_{i=1}^{n}\xi_i^{m+1} \\
\textbf{s.t.}\quad & y_i^l(w_l^T\phi_l(x_i^l)) \geq 1-\xi_i^l,\ 1\leq l\leq m,\ 1\leq i\leq n; \\
& \boldsymbol{d}\geq \mathbf{0},\quad \|\boldsymbol{d}\|_p^p \leq 1,\quad \boldsymbol{\xi}\geq \mathbf{0}, \\
& y_i^{m+1}\Big(\sum_{l=1}^{m+1} d_l(w_l+v_l)^T\phi_l(x_i^l)\Big) \geq 1-\xi_i^{m+1},\ \text{when } 1\leq i\leq n
\end{aligned} \tag{1}$$

where $\mathbf{0}$ is a column vector whose elements are all 0's. Inequalities between two vectors are taken element-wise, and $C_l, 1\leq l\leq m$ and C are positive user defined parameters.

To convert **P1** into a convex optimization problem, we can simply replace w_l , v_l $(1\leq l\leq m+1)$, and ξ_i^l $(1\leq i\leq n$ and $1\leq l\leq m)$ with $\hat{w}_l/d_l$, $\hat{v}_l/d_l$, and $\hat{\xi}_i^l/d_l$, respectively. If $d_l=0$, we define $a/d_l=\infty$ when $a\neq 0$, and $a/d_l=0$ when $a=0$. Omitting the *hat* notation for $\hat{w}$, $\hat{v}$, and $\hat{\xi}$ for simplicity, **P1** becomes

$$\textbf{P2}\quad \min_{\boldsymbol{w},\boldsymbol{v},\boldsymbol{\xi},\boldsymbol{d}} \frac{1}{2}\sum_{l=1}^{m+1}\frac{\|v_l\|^2}{d_l} + \frac{1}{2}\sum_{l=1}^{m}\frac{C_l\|w_l\|^2}{d_l} + \frac{1}{2}\sum_{l=1}^{m}\sum_{i=1}^{n}\frac{(\xi_i^l)^2}{d_l} + C\sum_{i=1}^{n}(\xi_i^{m+1}) \tag{2}$$

$$\textbf{s.t.}\quad y_i^l(w_l^T\phi_l(x_i^l)) \geq d_l-\xi_i^l,\ 1\leq l\leq m;\ 1\leq i\leq n \tag{3}$$

$$y_i^{m+1}\Big(\sum_{l=1}^{m+1}(w_l+v_l)\phi_l(x_i^l)\Big) \geq 1-\xi_i^{m+1},\ 1\leq i\leq n \tag{4}$$

$$\boldsymbol{d}\geq\mathbf{0},\qquad \|\boldsymbol{d}\|_p^p\leq 1,\qquad \boldsymbol{\xi}\geq\mathbf{0} \tag{5}$$

Now it is clear from **P2** that the quadratic error weighted by '$\boldsymbol{d}$' (see Eq. 1.d of the submitted paper) make it easy to formulate a convex problem. Without the '$\boldsymbol{d}$' in Eq. (1.d) of the submitted paper, the error terms for auxiliary tasks in **P2** become $\left(\xi_i^l\right)^2/d_l^2$ which is not convex. It is also important to note that analytically eliminating '$\boldsymbol{d}$' in **P2** is not so simple as that in MKL (nor its p-norm variant considered by Micchelli and Pontil (2007)) with the technique proposed by Rakotomamonjy et al. (2008), and Micchelli and Pontil (2007), because '$\boldsymbol{d}$' presents in both the objective function and auxiliary task constraints of **P2**.

We can obtain the dual of **P2** following a standard method as that used in the non-sparse MKL (Kloft et al., 2009). However, directly solving the dual problem may create some numerical problems as discussed by Kloft et al. (2009). Hence, we employ the cutting plane algorithm to solve **P2**. The dual objective will still be used for checking the stopping condition, i.e., the relative duality gap. Fixing $\boldsymbol{d}$, we can solve the partial Lagrangian w.r.t. $\boldsymbol{w},\boldsymbol{v}$, and $\boldsymbol{\xi}$. For brevity, we provide the semi-infinite programming formulation directly:

$$\begin{aligned}
\textbf{P3}\ \min_{\boldsymbol{d},\rho}\quad & \rho \qquad \textbf{s.t.} \qquad \boldsymbol{d}\geq\mathbf{0},\quad \|\boldsymbol{d}\|_p^p\leq 1, \\
& \textit{and}\quad \rho \geq \Big(d_1\boldsymbol{e}^T, d_2\boldsymbol{e}^T,\ldots,d_m\boldsymbol{e}^T,\boldsymbol{e}^T\Big)\boldsymbol{\alpha} - \frac{1}{2}\Big(\sum_{l=1}^{m+1} d_l Q_l\Big) \\
& \text{for all}\quad \boldsymbol{\alpha}\ \text{satisifying}\ \mathbf{0}\leq\boldsymbol{\alpha},\ \text{and}\ \boldsymbol{\alpha}^{m+1}\leq C
\end{aligned}$$

where $\boldsymbol{\alpha}$ is the Lagrange multiplier vector such that $\boldsymbol{\alpha}=\Big(\big(\boldsymbol{\alpha}^l\big)_{(l=1)}^{(m+1)}\Big)$, and for each l, $\boldsymbol{\alpha}^l=\Big(\big(\alpha_i^l\big)_{(i=1)}^{(n)}\Big)$, and

$$Q_l=\Big(\boldsymbol{\alpha}^l\Big)^T\left(\tfrac{1}{C_l}K_l\circ\Big(\boldsymbol{y}^l\big(\boldsymbol{y}^l\big)^T\Big)+\boldsymbol{I}\right)\boldsymbol{\alpha}^l+\overline{C}_l\Big(\boldsymbol{\alpha}^{m+1}\Big)^T\left(K_l\circ\Big(\boldsymbol{y}^{m+1}\big(\boldsymbol{y}^{m+1}\big)^T\Big)\right)\boldsymbol{\alpha}^{m+1}$$

$+\frac{2}{C_l}\left(\boldsymbol{\alpha}^l\right)^T\left(K_l\circ\left(\boldsymbol{y}^l\left(\boldsymbol{y}^{m+1}\right)^T\right)\right)\boldsymbol{\alpha}^{m+1}$, where $\overline{C}_l = 1/C_l + 1$, for $1 \le l \le m$.
When $l = m+1$,
$Q_{m+1} = \left(\boldsymbol{\alpha}^{m+1}\right)^T\left(K_{m+1}\circ\left(\boldsymbol{y}^{m+1}\left(\boldsymbol{y}^{m+1}\right)^T\right)\right)\boldsymbol{\alpha}^{m+1}$, where "$\circ$" denotes the element-wise product between two matrices.

P3 can be solved by the cutting plane algorithm with a standard QP (quadratic programming) to find its most violated constraint. For the *restricted master problem* (Sonnenburg et al., 2006) of this algorithm, we use `CVX`[1] directly, rather than solve an approximate problem (Kloft et al., 2009). Based on Slater's condition, we can use the relative duality gap as a stopping criterion and we set a threshold of 10^{-2} in our experiment.

2 Derivation for the Dual of P2

The dual of **P2** is used in checking the stopping condition. Suppose the Lagrangian multipliers for (3) and (4) in **P2** are $\left(\left(\alpha_i^l\right)_{(l,i)=(1,1)}^{(m+1,n)}\right)$. Multipliers for the three terms in (5) are $\left(\left(\gamma_l\right)_{l=1}^{m+1}\right)$, $(1/p)\beta$, and $\left(\left(\eta_i^l\right)_{(l,i)=(1,1)}^{(m+1,n)}\right)$ respectively.

Let's define $D = \{a/b : a > b, a \text{ is a positive even number, and } b \text{ is a positive odd number.}\}$

Proposition 1 *If $p \in D$, the Wolfe dual of **P2** is the following:*

$$\textbf{D1}\ \max_{\boldsymbol{\alpha}}\ \boldsymbol{e}^T\boldsymbol{\alpha}^{m+1} - \left(\sum_{l=1}^{m+1}\left(Q_l'\right)^{\frac{p}{p-1}}\right)^{\frac{p-1}{p}} \tag{6}$$

$$\textbf{s.t.}\ \ \mathbf{0} \le \boldsymbol{\alpha},\ \text{and}\ \boldsymbol{\alpha}^{m+1} \le C \tag{7}$$

for $1 \le l \le m$, we let

$$Q_l' = \max\left(\frac{1}{2}Q_l - \boldsymbol{e}^T\boldsymbol{\alpha}^l, 0\right) \tag{8}$$

where

$$\begin{aligned} Q_l =& \left(\boldsymbol{\alpha}^l\right)^T\left(\frac{1}{C_l}K_l\circ\left(\boldsymbol{y}^l\left(\boldsymbol{y}^l\right)^T\right) + \boldsymbol{I}\right)\boldsymbol{\alpha}^l \\ &+ \overline{C}_l\left(\boldsymbol{\alpha}^{m+1}\right)^T\left(K_l\circ\left(\boldsymbol{y}^{m+1}\left(\boldsymbol{y}^{m+1}\right)^T\right)\right)\boldsymbol{\alpha}^{m+1} \\ &+ \frac{2}{C_l}\left(\boldsymbol{\alpha}^l\right)^T\left(K_l\circ\left(\boldsymbol{y}^l\left(\boldsymbol{y}^{m+1}\right)^T\right)\right)\boldsymbol{\alpha}^{m+1} \end{aligned} \tag{9}$$

where $\overline{C}_l = 1/C_l + 1$ and "$\circ$" represents the element-wise product between two matrices. When $l = m+1$, $Q'_{m+1} = (1/2)Q_{m+1}$ where

$$Q_{m+1} = \left(\boldsymbol{\alpha}^{m+1}\right)^T\left(K_{m+1}\circ\left(\boldsymbol{y}^{m+1}\left(\boldsymbol{y}^{m+1}\right)^T\right)\right)\boldsymbol{\alpha}^{m+1}$$

[1] Available at: `http://stanford.edu/~boyd/cvx`

Proof. Taking derivative w.r.t. to the Lagrangian $\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \beta, \boldsymbol{\eta})$ of **P2** we have

$$\frac{\partial \mathcal{L}}{\partial w_l} = C_l \frac{w_l^T}{d_l} - \sum_{i=1}^{n} \alpha_i^l y_i^l \phi_l(x_i^l) - \sum_{i=1}^{n} \alpha_i^{m+1} y_i^{m+1} \phi_l(x_i^l), 1 \leq l \leq m \tag{10}$$

$$\frac{\partial \mathcal{L}}{\partial v_l} = \frac{v_l^T}{d_l} - \sum_{i=1}^{n} \alpha_i^{m+1} y_i^{m+1} \phi_l(x_i^l),\ 1 \leq l \leq m+1 \tag{11}$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i^l} = \frac{\xi_i^l}{d_l} - \alpha_i^l - \eta_i^l,\ 1 \leq l \leq m;\ 1 \leq i \leq n \tag{12}$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i^{m+1}} = C - \alpha_i^{m+1} - \eta_i^{m+1},\ 1 \leq i \leq n \tag{13}$$

$$\frac{\partial \mathcal{L}}{\partial d_l} = -\frac{C_l}{2} \frac{\|w_l\|^2}{d_l^2} - \frac{\|v_l\|^2}{d_l^2} - \frac{1}{2} \sum_{i=1}^{n} \frac{(\xi_i^l)^2}{d_l^2} + \boldsymbol{e}^T \boldsymbol{\alpha}^l + \beta d_l^{p-1} - \gamma_l,\ 1 \leq l \leq m \tag{14}$$

$$\frac{\partial \mathcal{L}}{\partial d_{m+1}} = -\frac{1}{2} \frac{\|v_{m+1}\|^2}{d_{m+1}^2} + \beta d_{m+1}^{p-1} - \gamma_{m+1}, \tag{15}$$

Setting all these to zero, and plugging all back into the Lagrangian, we have

$$\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \beta, \boldsymbol{\eta}) = \boldsymbol{e}^T \boldsymbol{\alpha}^{m+1} - \frac{1}{p}\beta - \frac{p-1}{p} \beta^{\frac{-1}{p-1}} \left(\sum_{l=1}^{m+1} \Big(G_l \Big)^{\frac{p}{p-1}} \right) \tag{16}$$

where for $1 \leq l \leq m$

$$G_l = Q_l'' + \gamma_l + \boldsymbol{\eta}_l^T \boldsymbol{\alpha}^l + \frac{1}{2} \boldsymbol{\eta}_l^T \boldsymbol{\eta}_l \tag{17}$$

and

$$G_{m+1} = Q_{m+1}'' + \gamma_{m+1} \tag{18}$$

where $Q_l'' = \frac{1}{2} Q_l - \boldsymbol{e}^T \boldsymbol{\alpha}^l$, $1 \leq l \leq m$ and $Q_{m+1}'' = Q_{m+1}'$.

We have $p \in D$ and this ensures the following: when solving for d_l from Eq.(14) after replacing $\|w_l\|^2/d_l^2, \|v_l\|^2/d_l^2$ and $(\xi_i^l)^2/d_l^2$ by Eq.(11), (12), and (13), we always obtain a real solution for d_l. Also, since $p \in D$, $\Big(G_l\Big)^{\frac{p}{p-1}}$ is always a non-negative value. Thus, to maximize $\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \beta, \boldsymbol{\eta})$: (a). if $Q_l'' \geq 0$, then γ_l and $\boldsymbol{\eta}_l = \mathbf{0}$; (b). if $Q_l'' \geq 0$, then $\gamma_l = -Q_l''$, and still $\boldsymbol{\eta}_l = \mathbf{0}$. Hence, to maximize Eq.(16), we can eliminate $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$, and replace all the Q_l'' with Q'. Further, following Kloft et al. (2009), let's take derivative of (16) w.r.t β and ignoring that it is non-negative. Now, at the maximum of the Lagrangian,

$$\beta = \left(\sum_{l=1}^{m+1} \Big(Q_l' \Big)^{\frac{p}{p-1}} \right)^{\frac{p-1}{p}} \tag{19}$$

and β here is always non-negative. This concludes the proof. ∎

References

M. Kloft, U. Brefeld, S. Sonnenburg, P. Laskov, K.R. Müller, , and A. Zien. Efficient and accurate lp-norm multiple kernel learning. In *Advances in Neural Information Processing Systems 22*, pp. 997–1005, 2009.

C.A. Micchelli, and M. Pontil. Feature space perspectives for learning the kernel. *Machine Learning*, 66(2): 297–319, 2007.

A. Rakotomamonjy, F. Bach, Y. Grandvalet, and S. Canu. SimpleMKL. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *The Journal of Machine Learning Research*, 7:1565, 2006.