Supplementary materials

1 Dual derivation of DCMSVM sub-problem

The sub-problem can be formulated in primal form as

$$\underset{w_1,\cdots,w_K,\xi_1,\cdots,\xi_N}{\text{minimize}} \quad \frac{\lambda}{2}\sum_{c=1}^{K} w_c^T w_c + \sum_{c=1}^{K}\alpha_c^T(w_c-\overline{w}_c) + \frac{\rho}{2}\sum_{c=1}^{K}\|w_c-\overline{w_c}\|^2 + \sum_{i=1}^{N}\xi_i, \tag{1}$$

$$\text{subject to} \quad (w_{y_i}-w_c)^T x_i \geq 1-\xi_i-\delta_{y_i,c} \quad \forall i=1,\cdots,N, c=1,\cdots,K \tag{2}$$

where $\delta_{y_i,c}=1$ if $y_i=c$, 0 otherwise. Notice for $c=y_i$ the inequality constraints become $\xi_i \geq 0$. Remove the constant terms in (1), we have the following equivalent problem.

$$\underset{w_1,\cdots,w_K,\xi_1,\cdots,\xi_N}{\text{minimize}} \quad \frac{\lambda+\rho}{2}\sum_{c=1}^{K} w_c^T w_c + \sum_{c=1}^{K}(\alpha_c-\rho\overline{w}_c)^T w_c + \sum_{i=1}^{N}\xi_i, \tag{3}$$

$$\text{subject to} \quad (w_{y_i}-w_c)^T x_i \geq 1-\xi_i-\delta_{y_i,c} \quad \forall i=1,\cdots,N, c=1,\cdots,K \tag{4}$$

We introduce multipliers μ for inequality constraints and form the Lagrangian.

$$L(w_1,\cdots,w_K,\xi_1,\cdots,\xi_N,\mu) = \frac{\lambda+\rho}{2}\sum_{c=1}^{K} w_c^T w_c + \sum_{c=1}^{K}(\alpha_c-\rho\overline{w}_c)^T w_c + \sum_{i=1}^{N}\xi_i - \sum_{i,c}\mu_{i,c}\left((w_{y_i}-w_c)^T x_i - 1 + \xi_i + \delta_{y_i,c}\right). \tag{5}$$

The dual function is

$$g(\mu) = \inf_{w_1,\cdots,w_K,\xi_1,\cdots,\xi_N} L(w_1,\cdots,w_N,\xi_1,\cdots,\xi_N,\mu). \tag{6}$$

Setting the derivatives of the Lagrangian with respect to w_c and ξ_i to zero, we get

$$\frac{\partial L}{\partial \xi_i} = 1-\sum_{c=1}^{K}\mu_{i,c} = 0 \quad \Rightarrow \sum_{c=1}^{K}\mu_{i,c} = 1. \tag{7}$$

Similarly,

$$\frac{\partial L}{\partial w_c} = (\lambda+\rho)w_c + \alpha_c - \rho\overline{w}_c - \left(- \sum_{i=1}^{N} \mu_{i,c} x_i + \sum_{i=1}^{N} \delta_{y_i,c} \left(\sum_{q=1}^{K} \mu_{i,q} \right) x_i \right) \tag{8}$$

$$= (\lambda+\rho)w_c + \alpha_c - \rho\overline{w}_c + \sum_{i=1}^{N} (\mu_{i,c} - \delta_{y_i,c}) x_i = 0, \tag{9}$$

which results in

$$w_c = \frac{1}{\lambda+\rho} \left(\rho\overline{w}_c - \alpha_c + \sum_{i=1}^{N} (\delta_{y_i,c} - \mu_{i,c}) x_i \right). \tag{10}$$

Substitute (7) into the Lagrangian, we obtain the dual function represented only using dual variables.

$$g(\mu) = \overbrace{\frac{\lambda+\rho}{2} \sum_{c=1}^{K} w_c^T w_c}^{S_3} + \overbrace{\sum_{c=1}^{K} \left(\alpha_c - \rho\overline{w}_c + \sum_{i=1}^{N} \mu_{i,c} x_i \right)^T w_c}^{S_1} - \overbrace{\sum_{i,c} \mu_{i,c} x_i^T w_{y_i}}^{S_2} - \sum_{i,c} \mu_{i,c} \delta_{y_i,c} + N \tag{11}$$

Next we substitute (10) into the dual objective function (11). The constant vector $\alpha_c - \rho\overline{w}_c$ is denoted by t_c.

$$S_1 = \frac{1}{\lambda+\rho} \sum_{c=1}^{K} \left(t_c + \sum_{i=1}^{N} \mu_{i,c} x_i \right)^T \left(\sum_{j=1}^{N} (\delta_{y_j,c} - \mu_{j,c}) x_j - t_c \right) \tag{12}$$

$$= \frac{1}{\lambda+\rho} \sum_{c=1}^{K} \left(\sum_{i,j} x_i^T x_j \mu_{i,c} (\delta_{y_j,c} - \mu_{j,c}) + \sum_{i=1}^{N} x_i^T t_c (\delta_{y_i,c} - 2\mu_{i,c}) - \|t_c\|^2 \right) \tag{13}$$

$$= \frac{1}{\lambda+\rho} \left(\sum_{i,j} x_i^T x_j \sum_{c=1}^{K} \mu_{i,c} (\delta_{y_j,c} - \mu_{j,c}) + \sum_{i=1}^{N} x_i^T (t_{y_i} - 2\sum_{c=1}^{K} t_c \mu_{i,c}) - \sum_{c=1}^{K} \|t_c\|^2 \right) \tag{14}$$

$$S_2 = \frac{1}{\lambda+\rho}\sum_{c,i}\mu_{i,c}x_i^T\left(\sum_{j=1}^{N}(\delta_{y_j,y_i}-\mu_{j,y_i})x_j - t_{y_i}\right) \tag{15}$$

$$= \frac{1}{\lambda+\rho}\left(\sum_{i,j}x_i^Tx_j\sum_{c=1}^{K}\mu_{i,c}(\delta_{y_j,y_i}-\mu_{j,y_i}) - \sum_{i=1}^{N}x_i^T\sum_{c=1}^{K}\mu_{i,c}t_{y_i}\right) \tag{16}$$

$$= \frac{1}{\lambda+\rho}\left(\sum_{i,j}x_i^Tx_j(\delta_{y_j,y_i}-\mu_{j,y_i}) - \sum_{i=1}^{N}x_i^Tt_{y_i}\right) \tag{17}$$

$$= \frac{1}{\lambda+\rho}\left(\sum_{i,j}x_i^Tx_j\sum_{c=1}^{K}\delta_{y_i,c}(\delta_{y_j,c}-\mu_{j,c}) - \sum_{i=1}^{N}x_i^Tt_{y_i}\right) \tag{18}$$

$$S_1 - S_2 = \frac{1}{\lambda+\rho}\left(-\sum_{i,j}x_i^Tx_j\sum_{c=1}^{K}(\delta_{y_i,c}-\mu_{i,c})(\delta_{y_j,c}-\mu_{j,c}) + 2\sum_{i=1}^{N}x_i^T(t_{y_i}-\sum_{c=1}^{K}t_c\mu_{i,c}) - \sum_{c=1}^{K}\|t_c\|^2\right) \tag{19}$$

$$S_3 = \frac{1}{2(\lambda+\rho)}\sum_{c=1}^{K}\left(\sum_{i=1}^{N}(\delta_{y_i,c}-\mu_{i,c})x_i - t_c\right)\left(\sum_{j=1}^{N}(\delta_{y_j,c}-\mu_{j,c})x_j - t_c\right) \tag{20}$$

$$= \frac{1}{\lambda+\rho}\left(\frac{1}{2}\sum_{i,j}x_i^Tx_j\sum_{c=1}^{K}(\delta_{y_i,c}-\mu_{i,c})(\delta_{y_j,c}-\mu_{j,c}) - \sum_{i=1}^{N}x_i^T(t_{y_i}-\sum_{c=1}^{K}t_c\mu_{i,c}) + \frac{1}{2}\sum_{c=1}^{K}\|t_c\|^2\right) \tag{21}$$

$$S_3 + S_1 - S_2 = \frac{1}{\lambda+\rho}\left(-\frac{1}{2}\sum_{i,j}x_i^Tx_j\sum_{c=1}^{K}(\delta_{y_i,c}-\mu_{i,c})(\delta_{y_j,c}-\mu_{j,c}) + \sum_{i=1}^{N}x_i^T(t_{y_i}-\sum_{c=1}^{K}t_c\mu_{i,c}) - \frac{1}{2}\sum_{c=1}^{K}\|t_c\|^2\right) \tag{22}$$

Substitue (22) into the dual objective function (11), we have the dual objective function

$$g(\mu) = \frac{1}{\lambda+\rho}\left(-\frac{1}{2}\sum_{i,j}x_i^Tx_j\sum_{c=1}^{K}(\delta_{y_i,c}-\mu_{i,c})(\delta_{y_j,c}-\mu_{j,c}) + \sum_{i=1}^{N}x_i^T(t_{y_i}-\sum_{c=1}^{K}t_c\mu_{i,c}) - \frac{1}{2}\sum_{c=1}^{K}\|t_c\|^2\right) - \sum_{i,c}\mu_{i,c}\delta_{y_i,c} + N \tag{23}$$

Finally, after removing the constants we have the dual problem

$$\underset{\mu}{\text{maximize}} \quad g(\mu) = -\frac{1}{2(\lambda+\rho)}\sum_{i,j}x_i^Tx_j\sum_{c=1}^{K}(\delta_{y_i,c}-\mu_{i,c})(\delta_{y_j,c}-\mu_{j,c}) - \sum_{i,c}\mu_{i,c}(\frac{x_i^Tt_c}{\lambda+\rho}+\delta_{y_i,c}), \tag{24}$$

$$\text{subject to} \quad \mu_{i,c}\geq 0, \quad \sum_{c=1}^{K}\mu_{i,c}=1, \quad \forall i=1,\cdots,N, \quad c=1,\cdots,K. \tag{25}$$

This problem is slightly different from the dual problem for Crammer&Singer[1] SVM formulation, where the coefficient for $\mu_{i,c}$ in the last term of the objective is just $\delta_{y_i,c}$.

2 Sequential dual method for the sub-problem

Let $C = \frac{1}{\lambda+\rho}$, $e_{i,c} = 1 - \delta_{y_i,c}$, $\beta_{i,c} = C(\delta_{y_i,c} - \mu_{i,c})$. Notice

$$\sum_{c=1}^{K} \mu_{i,c} = 1, \qquad \sum_{c=1}^{K} \delta_{y_i,c} = 1, \qquad \sum_{c=1}^{K} \beta_{i,c} = 0. \tag{26}$$

Also multiplying $g(\mu)$ by C and adding constant terms will not change the optimal solution. We can rewrite (24) and (25) as

$$\underset{\beta}{\text{maximize}} \quad h(\beta) = -\frac{1}{2} \sum_{i,j} x_i^T x_j \sum_{c=1}^{K} \beta_{i,c}\beta_{j,c} + \sum_{i,c} \beta_{i,c}(C x_i^T t_c - e_{i,c}), \tag{27}$$

$$\text{subject to} \quad \beta_{i,c} \leq C\delta_{y_i,c}, \quad \sum_{c=1}^{K} \beta_{i,c} = 0, \quad \forall i = 1, \cdots, N, \quad c = 1, \cdots, K. \tag{28}$$

Rewrite (10) as

$$w_c(\beta) = \sum_{i=1}^{N} \beta_{i,c} x_i - C t_c, \tag{29}$$

and put it in the dual formulation, which gives (we change the sign of the objective so maximization becomes minimization).

$$\underset{\beta}{\text{minimize}} \quad h(\beta) = \frac{1}{2} \sum_{c=1}^{K} \|w_c(\beta)\|^2 + \sum_{i,c} \beta_{i,c} e_{i,c}, \tag{30}$$

$$\text{subject to} \quad \beta_{i,c} \leq C\delta_{y_i,c}, \quad \sum_{c=1}^{K} \beta_{i,c} = 0, \quad \forall i = 1, \cdots, N, \quad c = 1, \cdots, K. \tag{31}$$

The gradient of h is given by

$$h_i^c = \frac{\partial h(\beta)}{\partial \beta_{i,c}} = w_c(\beta)^T x_i + e_{i,c}, \quad \forall i = 1, \cdots, N, \quad c = 1, \cdots, K. \tag{32}$$

Now our ADMM sub–problem has been reduced to a form very close to what Keerthi *et al* used in their paper [2]. In fact the only difference between the two is that we have an extra constant term Ct_c for each $w_c(\beta)$. Given that these terms are independent of βs, and w is incrementally updated, if we initialize Keerthi *et al*'s algorithm 2.1 with $\beta_{i,c} = 0$ (or $\alpha = 0$ by their notation) and with $w_c = -Ct_c$, the solution it gives will be the solution to our ADMM sub–problem.
Based on this reduction, with a little modification to the part of code for solving the Crammer&Singer SVM, LibLinear package is ready to solve our ADMM sub–problem.

References

[1] *K.Crammer and Y.Singer, On the Algorithmic Implementation of Multiclass Kernel–based Vector Machines, Journal of Machine Learning Research 2 (2001) 265-292.*

[2] *S. S. Keerthi, S. Sundararajan, K.-W. Chang, C.-J. Hsieh, and C.-J. Lin. A sequential dual method for large scale multi-class linear SVMs. In ACM KDD, 2008.*