Due to numerical instability, when writing a computer program it is preferable to compute the lists of indices I_λ and I_μ given by

$$\begin{aligned} I_\lambda &= \{i \in \{1, \ldots, p\} \mid 0 < \lambda_i < K\} \\ I_\mu &= \{j \in \{1, \ldots, q\} \mid 0 < \mu_j < K\}. \end{aligned}$$

Then it is easy to see that we can compute b using the following averaging formula

$$b = w^\top \left(\Big(\sum_{i \in I_\lambda} u_i \Big) / |I_\lambda| + \Big(\sum_{j \in I_\mu} v_j \Big) / |I_\mu| \right) / 2.$$

Recall that $\delta = 1/\|w\|$.

Remark: There is a cheap version of Problem (SVM$_{s2}$) which consists in dropping the term $(1/2)w^\top w$ from the objective function:

Soft margin classifier (SVM$_{s2l}$):

$$\begin{aligned} &\text{minimize} \quad \sum_{i=1}^{p} \epsilon_i + \sum_{j=1}^{q} \xi_j \\ &\text{subject to} \\ &\qquad w^\top u_i - b \geq 1 - \epsilon_i, \quad \epsilon_i \geq 0 \qquad i = 1, \ldots, p \\ &\qquad - w^\top v_j + b \geq 1 - \xi_j, \quad \xi_j \geq 0 \qquad j = 1, \ldots, q. \end{aligned}$$

The above program is a linear program that minimizes the number of misclassified points but does not care about enforcing a minimum margin. An example of its use is given in Boyd and Vandenberghe; see [29], Section 8.6.1.

The "kernelized" version of Problem (SVM$_{s2}$) is the following:

Soft margin kernel SVM (SVM$_{s2}$):

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\langle w, w\rangle + K \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q} \\ &\text{subject to} \\ &\qquad \langle w, \varphi(u_i)\rangle - b \geq 1 - \epsilon_i, \quad \epsilon_i \geq 0 \qquad i = 1, \ldots, p \\ &\qquad - \langle w, \varphi(v_j)\rangle + b \geq 1 - \xi_j, \quad \xi_j \geq 0 \qquad j = 1, \ldots, q. \end{aligned}$$

Redoing the computation of the dual function, we find that the dual program is given by