of the values of the primal and dual objective functions. Yet others stop simply when no further improvement is made [398].

Before we develop a stopping criterion recall that ultimately we want to find a solution $f(x) = \langle \mathbf{w}, \Phi(x) \rangle + b$ that minimizes one of the regularized risk functionals described below. In the case of classification,

$$\begin{array}{lll} \underset{\mathbf{w},\xi}{\text{minimize}} & C\sum_{i=1}^{m} c(\xi_i) + \frac{1}{2}\|\mathbf{w}\|^2 & \sum_{i=1}^{m} c(\xi_i) + \frac{1}{2}\|\mathbf{w}\|^2 - \nu\rho \\ \text{subject to} & y_i f(x_i) \geq 1 - \xi_i \quad \text{or} & y_i f(x_i) \geq \rho - \xi_i \\ & \xi_i \geq 0 & \xi_i \geq 0, \rho \in \mathbb{R} \end{array} \tag{10.3}$$

(the right half of the equations describes the analogous setting with the ν-parameter), similarly, for regression,

$$\begin{array}{lll} \underset{\mathbf{w},\xi}{\text{minimize}} & C\sum_{i=1}^{m} c(\xi_i) + c(\xi_i^*) + \frac{1}{2}\|\mathbf{w}\|^2 & C\sum_{i=1}^{m} c(\xi_i) + c(\xi_i^*) + \frac{1}{2}\|\mathbf{w}\|^2 - \nu\epsilon \\ \text{subject to} & f(x_i) \geq y_i - \epsilon - \xi_i \quad \text{or} & f(x_i) \geq y_i - \epsilon - \xi_i \\ & f(x_i) \leq y_i + \epsilon + \xi_i^* & f(x_i) \leq y_i + \epsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0 & \xi_i, \xi_i^* \geq 0, \epsilon \in \mathbb{R} \end{array} \tag{10.4}$$

This means that ultimately not the Lagrange multipliers α_i but rather $\mathbf{w}$, or only the value of the primal objective function, matters. Thus, algorithms [266, 290, 291, 398] which rely on the assumption that proximity to the optimal parameters will ensure a good solution may not be using an optimal stopping criterion. In particular, such a criterion may sometimes be overly conservative, especially if the influence of individual parameters on the final estimate is negligible. For instance, assume that we have a linear dependency in the dual objective function. Then there exists a linear subspace of parameters which would all be suitable solutions, leading to identical vectors $\mathbf{w}$. Therefore, convergence within this subspace may not occur and, even if it does, it would not be relevant to the quality of the solution.

Proximity in Parameters ≠ Proximity in Solution

What we would prefer to have is a way of bounding the distance between the objective function at the current solution f and at f_{opt}. Since (10.3) and (10.4) are both constrained optimization problems we may make use of Theorem 6.27 and lower bound the values of (10.3) and (10.4) via the KKT Gap. The advantage is that we do not have to know the optimal value in order to assess the quality of the approximate solution. The following Proposition formalizes this connection.

Proposition 10.1 (KKT-Gap for Support Vector Machines) *Denote by f the (possibly not optimal) estimate obtained during a minimizing procedure of the optimization problem (10.3) or (10.4) derived from the regularized risk functional $R_{\text{reg}}[f]$. Further, denote by f* the minimizer of $R_{\text{reg}}[f]$. Then under the condition of dual feasible variables (namely that the equality and box constraints are satisfied), the following inequality holds:*

$$R_{\text{reg}}[f] \geq R_{\text{reg}}[f^*] \geq R_{\text{reg}}[f] - \frac{1}{Cm}\text{Gap}[f] \tag{10.5}$$

where $\text{Gap}[f]$ *is defined as follows:*