1 Optimization of SVM

A: Convex Optimization: multiplicity of solutions in SVM The variables to the dual SVM optimization are the Lagrange parameters α_i, with one Lagrange parameter per datapoint, i.e. $i = 1...M$. As per the KKT conditions, the Lagrange parameters represent the weight given to each datapoints to construct $w = \sum_i \alpha_i x^i$.

Can we find different sets of α_i that lead to the same optimum?

Let $w = \alpha_1 x^1 + \alpha_2 x^2$ be the optimal w. Since none of the datapoints are collinear, any pair of two points is linearly independent. Hence, each point can be expressed as linear combination of two other points.

We can hence construct $x^2 = \beta_1 x^1 + \beta_2 x^3$ with appropriate scalars β_1, β_2. Replacing x^2 in w, we obtain a new set of α_i for the same optimal w, namely $w = (\alpha_1 + \beta_1)x^1 + \beta_2 x^3$.

B: Margin The KKT condition $\sum_i \alpha_i y_i = 0$ implies that we have at least two support vectors, one in each class. Hence, there exist two points , which we denote as x^1 and x^2 with $y_1 = 1$ and $y_2 = -1$, for which the constraints $y_i(w^T x^i + b) = 1$ are satisfied.

We modify the constraint and set that all support vectors lie on a plane with equation $y_i(w^T x^i + b) = a$, with $a > 0$. We have:

$$\begin{cases} w^T x^1 + b = a \\ w^T x^2 + b = -a \end{cases} \tag{1}$$

Substracting the two lines, we get $w^T(x^2 - x^1) = 2a$. Expanding the inner product, $\|w\| = \frac{2a}{\|(x^2 - x^1)\| \cos(\theta)}$. θ is the angle between w and the vector $x^2 - x^1$. We see that the factor a only scales the norm of the vector w, but does not affect the choice of Support Vectors. It does not change the direction of w and hence does not affect the orientation of the hyperplane.

C: Convexity of the relaxed problem Is $f(w, \xi) = \|w\|^2 + C \sum_i \xi_i$, $\xi_i > 0 \forall i, C \geq 0$ convex?

$f(w, \xi) = \|w\|^2$ is strictly convex and $\sum_i \xi$, $\xi > 0$, $\forall i$ is convex. Since the quadratic term is strictly convex and grows faster than the linear term, the objective function is strictly convex. It hence admits a single global optimum.

The addition of the slack variables, however, can shift the optimum of the objective function to a solution that is not the true optimum (without relaxation of constraints). The relaxed optimization finds an optimal solution that is a tradeoff between augmenting the margin across the two classes (reducing the first term of the cost function) and reducing the cost of violating one or more constraints (reducing the second term of the cost function).

The penalty associated to the violation of the constraint is conveyed through the choice of the constant C. A large C will tend to force the optimization to find a solution close to the unrelaxed problem. This is illustrated in Figure 1. When applying a small penalty, $C = 5$, for a violation of the constraints, the optimization finds a separating hyperplane with a larger margin than with a hight penalty, $C = 100$.

D: Optimum of the relaxed problem: The true optimal solution to SVM is obtained for an optimal value to the objective function and satisfaction of all constraints. In the relaxed problem, the objective function is given by: $min_{w,\xi} \|w\|^2 + \frac{C}{M} \sum_i \xi_i$, with $C \geq 0$ a constant penalizing for the introduction of slacks and M the number of datapoints. Observe that the SVM objective function is composed of a quadratic and linear cost, both of which are proportional to the width of the margin, which we denote as a.

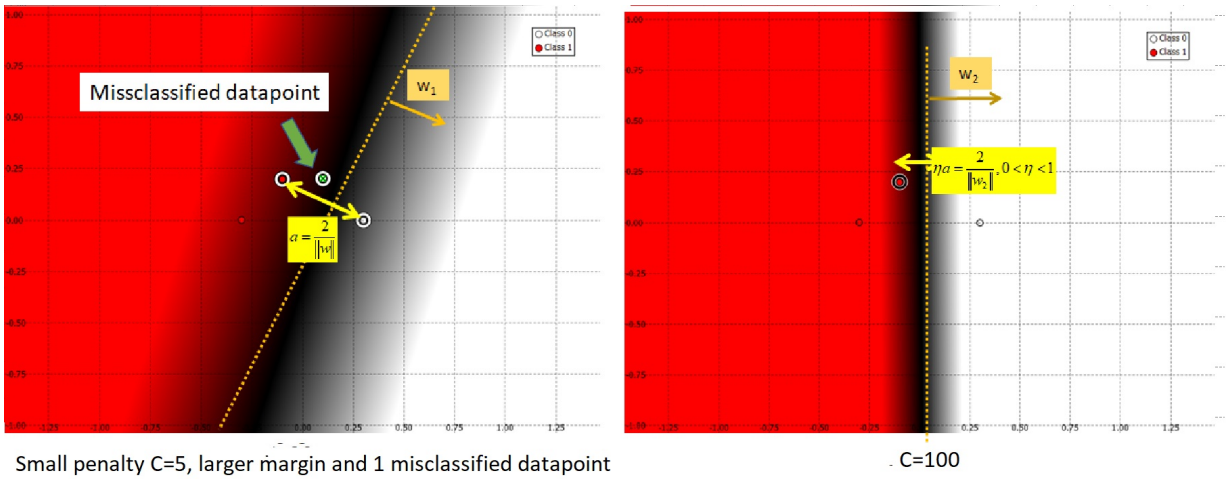

Figure 1: Optimal solution of the relaxed SVM optimization when using a low penalty on slacks $C = 5$ versus a high penalty, $C = 100$.

Consider the group of four points in Figure 1. The two hyperplanes generated by w_1 amd w_2, both optimal solutions for different values of C.

The first hyperplane defined by w_1 has a margin equal to $\|w_1\|^2 = \frac{2}{a^2}$. One of the two points from the white class is missclassfied. The costs associated to the constraint's violation for this point is entailed in the associated slack ξ. We show next that the slack is proportional to the distance to the hyperplane.

Without loss of generality, we can assume $b = 0$ (shift of the origin). The constraints are satisfied at equality for the two datapoints on the margin and for the point inside the margin with slack ξ. For the latter, we have:

$$\begin{cases} w_1^T x^i = 1 + x, \\ \xi = \|w_1\|\|x\| - 1. \end{cases} \tag{2}$$

The second hyperplane w_2 satisfies all constraints, hence $\xi = 0, \forall i$ and is solution to $\|w_2\|^2 + C\sum_i \xi_i = \|w\|^2 = \frac{2}{(\eta a)^2}$, $0 < \eta \leq 1$.

To determine if a solution with slack can lead to a value on the objective function that is equal or better than the solution without slack, one must hence check whether $\|w_1\|^2 + C\xi = \frac{2}{a^2} + C\frac{1}{\eta a} \leq \|w_2\|^2 = \frac{2}{(\eta a)^2}$. Many cases will arise depending on the values of C and η. Observe that the associated cost on the objective function to enlarging the margin is privileged over violating constraints, as the former grows quadratically with the margin whereas the latter grows linearly. The solver will hence tend to privilege solutions with small violation of constraints if these lead to an increase in the margin. The shift of the optimum is illustrated in Figure 2.

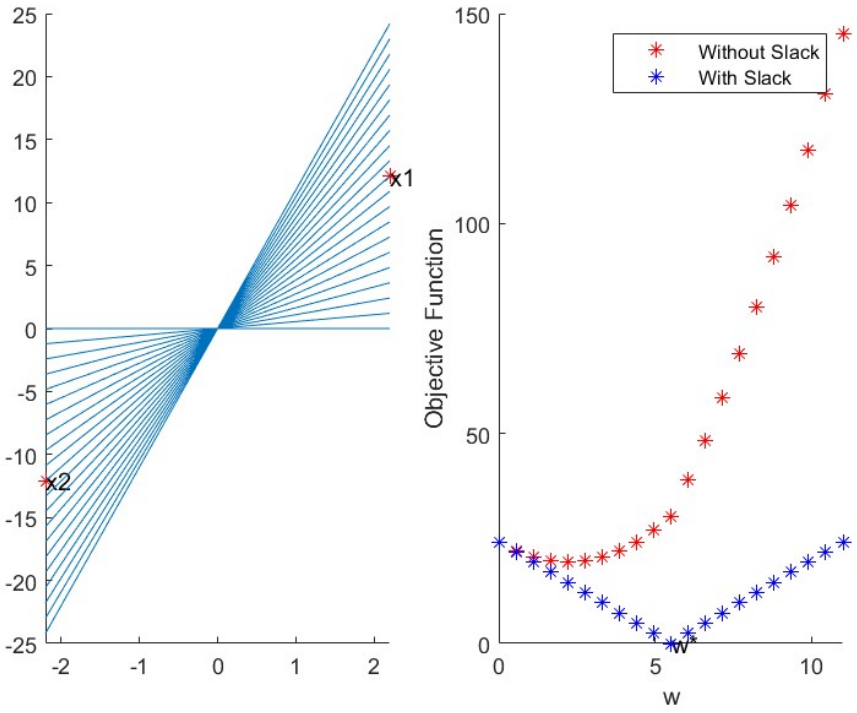

Figure 2: (Left) distribution of separating hyperplanes across a pair of datapoint. (Right) evolution of optimum on SVM objective function for the distribution of hyperplane with and without slack.