CSC 446
Lecture Notes
SUPPORT VECTORS

prepared by Phyo Thiha

(Note: Please read more about the Wolfe dual in a separate note shared on class website at http://www.cs.rochester.edu/u/gildea/2012_Spring/wolfe.pdf)
According to Wolfe duality, we have

$$\min f_0(x) \text{ s.t. } f_i(x) \leq 0 \text{ for } i=1...I \qquad \text{----------(1)}$$

Suppose we take Lagrangian of equation (1),

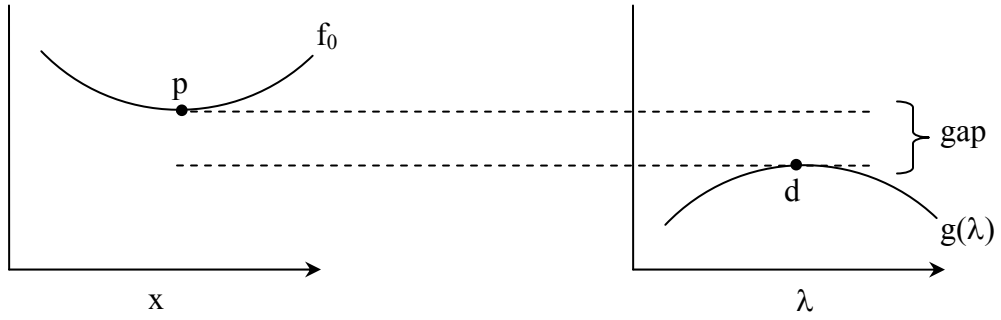
$$g(\lambda) = L(x, \lambda) = f_0(x) + \Sigma \lambda_i f_i(x)$$

$$\max\_\lambda g(\lambda) \text{ s.t. } \lambda_i \geq 0 \qquad \text{----------(2)}$$

Equation (1) is called primal problem and (2) is called dual. If we solve this duality, it's always true that

$$g(\lambda) \leq f_0(x)$$

for $\forall x$ where 'λ' is feasible. This relationship is called **weak duality**. To visualize this relationship-



In gradient descent problem, how do we know when to stop? One way to test that is to find a point in dual problem. There, we know that 'p' > 'd' and we can see the gap as how much room there is for improvement in our gradient descent.

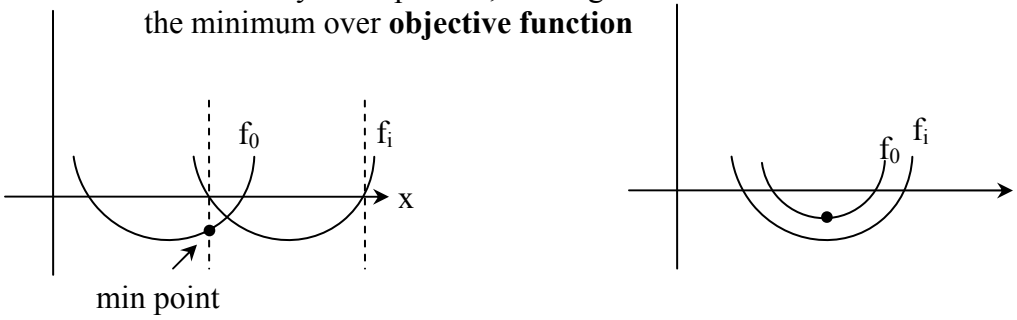
In SVM, we will utilize the above knowledge. We will morph the first problem (primal) into second form (dual) and solve the second one. Then we will map it back to the original problem. We will also assume strong duality (where d = p)

Note: Why did we choose the dual instead of the primal? Because in dual, we're dealing with 'λ', which is not dependent on anything else whereas in primal, we have $f_i(x)$ constraint to deal with (therefore, more complex)

Above plan of utilizing duality to solve SVM works ONLY IF

1) f_0 and f_i are convex (this is a convex optimization problem)

convex – if you drop a ball, it will go to the minimum over **objective function**



2) f_0 and f_i are differentiable
3) feasible set has an interior (Slater's condition)

If all these assumptions are true, then strong duality holds. Suppose we have found,

$$d^* = g(\lambda^*)$$

Case 1: Suppose $\lambda_i = 0$ (meaning the bound is tight). This means

$$\frac{\partial L}{\partial \lambda_i} \leq 0 \qquad \text{----------(2)}$$

$$g(\lambda) = \min\_x\ L(x, \lambda)$$

$$\frac{\partial g}{\partial \lambda} < 0$$

$$g(0) = L(x', 0) \qquad \text{----------(3)}$$

From (2) and (3), $f_i(x') < 0$ [because $\frac{\partial L}{\partial \lambda_i} = f_i$]

$$\frac{\partial L}{\partial x}|_{x'} = 0 \text{ (we get 0 at x')} \qquad \text{----------(4)}$$

$$\frac{\partial L}{\partial x} = \nabla f_0(x) + \sum_i \lambda_i \nabla f_i(x) \qquad \text{----------(5)}$$

From (4) and (5),

$$0 = \nabla f_0(x) + \sum_i \lambda_i \nabla f_i(x)$$

$$0 = \nabla f_0(x) \qquad (\sum_i \lambda_i \nabla f_i(x) \text{ is zero}) \quad \text{----------(6)}$$

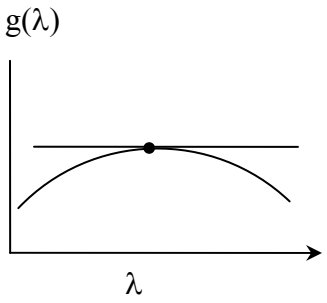
(6) tells us that we are at the minimum value in the constraint. Therefore,

$$x' = x^*$$
$$p^* = f_0(x^*) = d^*$$

That is, max of 'g' is min of 'L' and we satisfy the strong duality. END of case $\lambda_i = 0$.

==========

Case 2: $\lambda_i > 0$



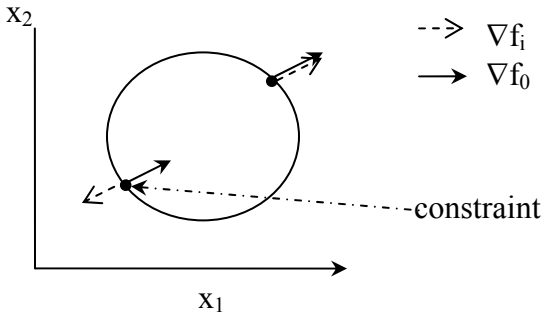
$$\frac{\partial g}{\partial \lambda} = 0$$

$x' = \text{argmin\_x } L(x, \lambda_i^*)$ [in other words, x' is the 'x' that gives us the min. point]

$$\frac{\partial L}{\partial \lambda_i} = f_i(x) = 0$$

$$\frac{\partial L}{\partial x}|_{x'} = 0 = \nabla f_0(x) + \sum_i \lambda_i \nabla f_i(x)$$

$$g(\lambda^*) = L(x', \lambda^*) = f_0(x') + f_i(x')$$



There are more complex conditions where two constraints interplay (called Slater's condition), but we won't go into that.
= = = = = = = = =

Karush Kuhn Tucker (KKT) conditions

1) $\frac{\partial L}{\partial x} = 0$
2) $f_i(x) \leq 0$ (feasible) and
3) $\lambda_i \geq 0$ (feasible)
4) $\lambda_I f_i(x) = 0$ (complementary slack)

If we can find $L(x, \lambda)$ such that it satisfies above conditions, we have solved the optimization problem.
= = = = = = = = = =

We will use KKT for solving SVM. We want

$$\max \frac{1}{2}\|w\|^2 \qquad \text{s.t.} \qquad y^n(w^T x^n) \geq 1$$

where 'y^n' represents each data point in the training and can be ± 1.

We'll add the constant (bias?) back in. That is pretty much equivalent to saying that it's okay for some labels to be on the wrong side of the boundary (esp. in cases where data is linearly inseparable).

For each data point which is on the wrong side of the boundary, we will penalize for it.

$$\max \frac{1}{2}\|w\|^2 + c\sum_n \xi_n$$

where 'c' is capacity and 'ξ' is the penalizing term (should always be positive).

$$\xi_n \geq 0 \qquad \rightarrow \mu_n$$

$$y^n(w^T x^n + b) + \xi_n \geq 1 \qquad \rightarrow \alpha_n$$

Let's solve:

$$L(x, \xi) = L(w, b, \xi, \alpha, \mu)$$
$$= \frac{1}{2} w^T w + c\sum_n \xi_n - \sum_n \alpha_n (y^n (w^T x^n + b) + \xi_n - 1) - \sum_n \mu_n \xi_n \quad \text{-----(7)}$$

$\frac{\partial L}{\partial w} = w - \sum_n \alpha_n y^n x^n = 0$, which leads to

$$\sum_n \alpha_n y^n x^n = w$$

$$\frac{\partial L}{\partial b} = -\sum_n \alpha_n y^n = 0 \text{ [how did we drop } y^n \text{ here?]}$$

$$\frac{\partial L}{\partial \xi_n} = c - \alpha_n - \mu_n = 0 \text{, which leads to}$$

$$\mu_n = c - \alpha_n$$

Since $\mu_n \geq 0$,

$$c - \alpha_n \geq 0$$

$$c \geq \alpha_n \qquad \text{----------(8)}$$

Substitute μ_n in equation (7),

$$L = \frac{1}{2} w^T w + c\sum_n \xi_n - \sum_n \alpha_n (y^n (w^T x^n + b) + \xi_n - 1) - \sum_n \mu_n \xi_n$$

$$= \frac{1}{2}\left(\sum_n \alpha_n y^n x^n\right)^T \left(\sum_n \alpha_n y^n x^n\right) - \sum_n \alpha_n \left(y^n \left(\left(\sum \alpha_n y^n x^n\right)^T x^n + b \right) - 1 \right)$$

$$= \frac{1}{2}\left\|\sum_n \alpha_n y^n x^n\right\|^2 - \sum_n \alpha_n y^n b - \sum_n \alpha_n$$

$$= \frac{1}{2}\sum_n \sum_m \alpha_n \alpha_m y^n y^m x^{n^T} x^m - \sum_n \alpha_n$$

From equation (8), max_α $g(\alpha)$ s.t. $\alpha_n \geq 0$, $\alpha_n \leq c$ (this gives us a boundary to focus our optimization problem on instead of trying for all possible space.

[Disclaimer/Apology: I was not able to get some of the jumps in calculation that Dan assumed in class. Also, I'm still shaky myself w.r.t. some of the concepts/assumptions that we made before calculating SVM. It'd be the best if you ask Dan or read some more explanation online].