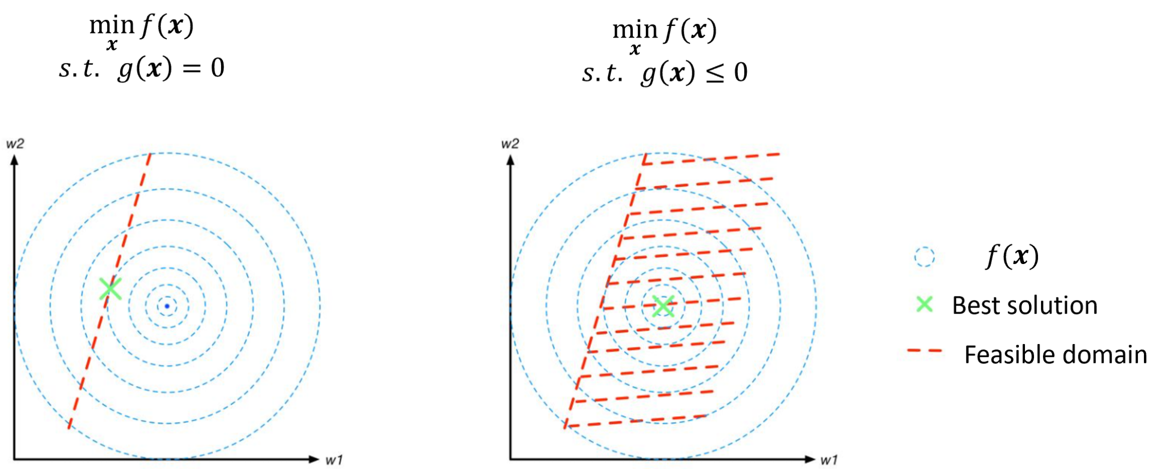
Understand dual problem and KKT condition

Equality and inequality constraint problem



Equality condition

$\boldsymbol{x}^*$ is where the best solution is , and we could easily get:

$\nabla f\left(\boldsymbol{x}^*\right) \perp g(\boldsymbol{x})=0$ and $\nabla g\left(\boldsymbol{x}^*\right) \perp g(\boldsymbol{x})=0$

######

$$\Rightarrow \nabla f\left(\boldsymbol{x}^*\right)+\lambda \nabla g\left(\boldsymbol{x}^*\right)=0 \tag{1.1}$$

so, we could write this Lagrangian equation

$$L(\boldsymbol{x}, \lambda)=f(\boldsymbol{x})+\lambda g(\boldsymbol{x}) \tag{1.2}$$

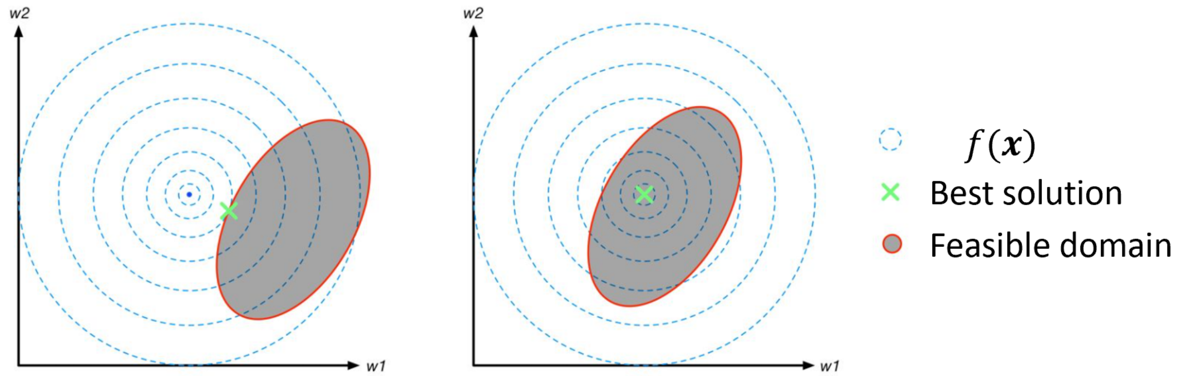
if you try to do partial guidance for $\boldsymbol{x}$ and λ , you could find it's just (1.2) and $g(\boldsymbol{x})=0$

so we could say $\min_{\boldsymbol{x}} L$ is same as the equality condition

But how to deal with the inequality condition?

KKT condition

There are two possible condition for the inequality constraint problem:



The left one means $\boldsymbol{x}^*$ is at the boundary but the best x for $f(\boldsymbol{x})$ is out side the feasible domain, so the direction of $\nabla f(\boldsymbol{x}^*)$ is opposite to the direction of $\nabla g(\boldsymbol{x}^*)$, so we could tell from the (1.1) that $\lambda < 0$

The right conditions shows that $\boldsymbol{x}^*$ is in the domain and the best x for $f(\boldsymbol{x})$ is just the $\boldsymbol{x}^*$ so the constraint equation is not working so that $\lambda = 0$

Based on that , we could now get our KKT condition:

$$\begin{gathered} g(\boldsymbol{x}) \leqslant 0 \\ \lambda \geqslant 0 \\ \lambda g(\boldsymbol{x}) = 0 \end{gathered} \tag{1.3}$$

Lagrangian dual problem

Target function

Let’s back to a trivial problem

$$\begin{array}{ll} \min_{\boldsymbol{x}} & f(\boldsymbol{x}) \\ \text{s.t.} & h_i(\boldsymbol{x}) = 0 \quad i = 1, 2, \ldots, m \\ & g_j(\boldsymbol{x}) \leq 0 \quad j = 1, 2, \ldots, n \end{array} \tag{2.1}$$

2.1 shows multiple constraint problems (m equality and n equality)

Lagrangian equation

The L equation of (2.1) is :

$$L(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \alpha_i h_i(\boldsymbol{x}) + \sum_{j=1}^{n} \beta_j g_j(\boldsymbol{x}) \tag{2.2}$$

Our goal is to solve (2.1), why we need (2.2) is a minmax problem.

Firstly,we build a new target function(primary equation):

$$\theta_P(\boldsymbol{x}) = \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}; \beta_j \geq 0} L(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\boldsymbol{x}) + \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}; \beta_j \geq 0} \left[\sum_{i=1}^{m} \alpha_i h_i(\boldsymbol{x}) + \sum_{j=1}^{n} \beta_j g_j(\boldsymbol{x}) \right] \tag{2.3}$$

Write this feasible domain as Φ, so inside the domain when $\beta_j \geq 0, j = 1, 2, \ldots, n$, we got:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta};\beta_j\geq 0}\left[\sum_{i=1}^{m}\alpha_i h_i(\boldsymbol{x})+\sum_{j=1}^{n}\beta_j g_j(\boldsymbol{x})\right]=0,\ \text{for } \boldsymbol{x}\in\Phi \tag{2.4}$$

when it's outside the domain, it means at least one set of the constraint in (2.1) is not met. if $h_i(\boldsymbol{x})\neq 0$, we could let $\alpha_i h_i(\boldsymbol{x})\rightarrow+\infty$, if $g_j(\boldsymbol{x})>0$, we could let $\beta_j g_j(\boldsymbol{x})\rightarrow+\infty$, so now:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta};\beta_j\geq 0}\left[\sum_{i=1}^{m}\alpha_i h_i(\boldsymbol{x})+\sum_{j=1}^{n}\beta_j g_j(\boldsymbol{x})\right]=+\infty,\ \text{for } \boldsymbol{x}\notin\Phi \tag{2.5}$$

Combine (2.3),(2.4) and (2.5) we got :

$$\theta_P(\boldsymbol{x})=\begin{cases} f(\boldsymbol{x}) & \boldsymbol{x}\in\Phi \\ +\infty & \text{otherwise}\end{cases} \tag{2.6}$$

Now we could use :

$$\min_{x}\left[\theta_P(\boldsymbol{x})\right] \tag{2.7}$$

to solve the 2.1!

Dual problem

It's hard to solve:

$$\min_{\boldsymbol{x}}\left[\theta_P(\boldsymbol{x})\right]=\min_{\boldsymbol{x}}\left[\max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\beta_j\geq 0}L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})\right] \tag{3.1}$$

because we need to remove α and β

so we have to consider it's dual problem:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\beta_j\geq 0}\left[\theta_D(\boldsymbol{\alpha},\boldsymbol{\beta})\right]=\max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\beta_j\geq 0}\left[\min_{\boldsymbol{x}}L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})\right] \tag{3.2}$$

It could be proved that the solution of (3.2) and (3.1) are exactly the same when it meets KKT condition!

The detailed proof is here(*Convex Optimization* by Boyd and Vandenberghe. Page-234, 5)https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf

Back to SVM

The problem we are going to solve is :

$$\begin{aligned} &\min_{\boldsymbol{w},b}\ \tfrac{1}{2}\|\boldsymbol{w}\|^2 \\ &\text{s.t. } y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b\right)\geqslant 1,\quad i=1,2,\ldots,m\end{aligned} \tag{4.1}$$

Write the Langrangian equation of (4.1):

$$L(\boldsymbol{w},b,\boldsymbol{\alpha})=\frac{1}{2}\|\boldsymbol{w}\|^2+\sum_{i=1}^{m}\alpha_i\left(1-y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b\right)\right) \tag{4.2}$$

first calculate the inside part of (3.2) which is to calculate the partial derivative for $\boldsymbol{w},b$ we could get (please notice here x is not the independent variable , they are the sample points!)

$$\begin{aligned} \boldsymbol{w}&=\textstyle\sum_{i=1}^{m}\alpha_i y_i\boldsymbol{x_i} \\ 0&=\textstyle\sum_{i=1}^{m}\alpha_i y_i\end{aligned} \tag{4.3}$$

Put (4.3) in (4.2)

we got

$$
\begin{aligned}
\max_{\alpha} & \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x_i}^{\mathrm{T}} \boldsymbol{x_j} \\
\text{s.t.} \quad & \sum_{i=1}^{m} \alpha_i y_i = 0 \\
& \alpha_i \geqslant 0, \quad i = 1, 2, \ldots, m
\end{aligned} \tag{4.4}
$$

it also need to meet the KKT condition:

$$
\left\{\begin{array}{l}
\alpha_i \geqslant 0 \\
y_i f\left(\boldsymbol{x}_i\right) - 1 \geqslant 0 \\
\alpha_i \left(y_i f\left(\boldsymbol{x}_i\right) - 1\right) = 0
\end{array}\right. \tag{4.5}
$$

Please notice that in the calculation, we use $\boldsymbol{x_i}^{\mathrm{T}} \boldsymbol{x_j}$ in the calculation!

In the high dimension space,$\boldsymbol{x} \mapsto \phi(\boldsymbol{x})$, we write $\kappa\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \left\langle \phi\left(\boldsymbol{x}_i\right), \phi\left(\boldsymbol{x}_j\right) \right\rangle = \phi(\boldsymbol{x}_i)^{\mathrm{T}} \phi\left(\boldsymbol{x}_j\right)$

so that we don’t need to know what $\phi(\boldsymbol{x})$ is