# Support Vector Regression

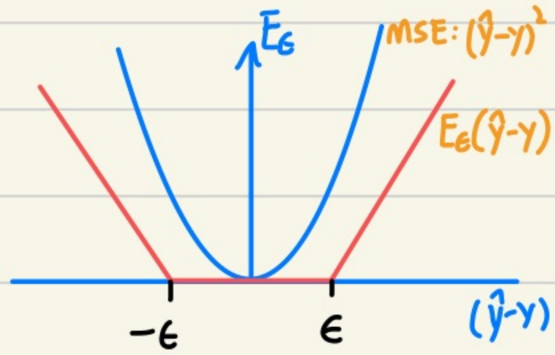
In Ridge Regression $J(\underline{w}) = \frac{1}{N}\|\underline{X}\,\underline{w} - \underline{y}\|_2^2 + \lambda\|\underline{w}\|_2^2$

or $\frac{1}{N}\sum_{n=1}^{N}[\hat{y}(x_n) - y_n]^2 + \lambda\|\underline{w}\|_2^2 \quad (\lambda > 0)$

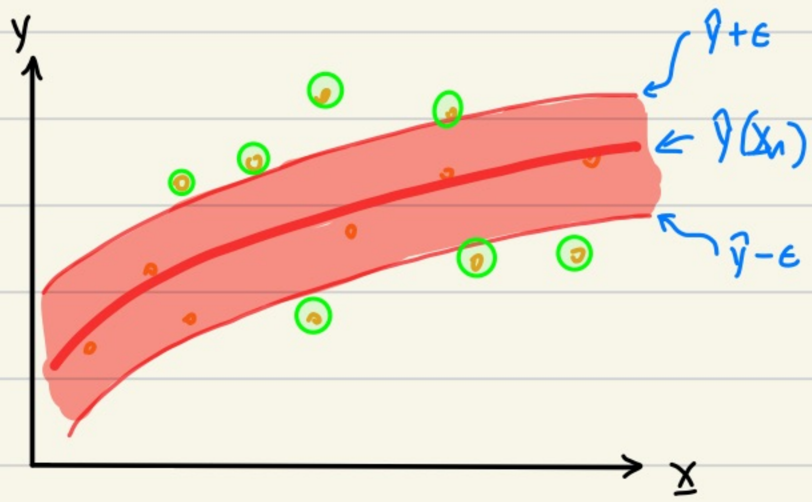
In SVR $J(\underline{w}) = C \cdot \sum_{n=1}^{N} E_\epsilon[\hat{y}(x_n) - y_n] + \frac{1}{2}\|\underline{w}\|_2^2$

by convention

where $E_\epsilon(\hat{y} - y) = \begin{cases} 0 & \text{if } |\hat{y} - y| < \epsilon \\ |\hat{y} - y| - \epsilon & \text{if } |\hat{y} - y| \geq \epsilon \end{cases}$




Visual SVR:



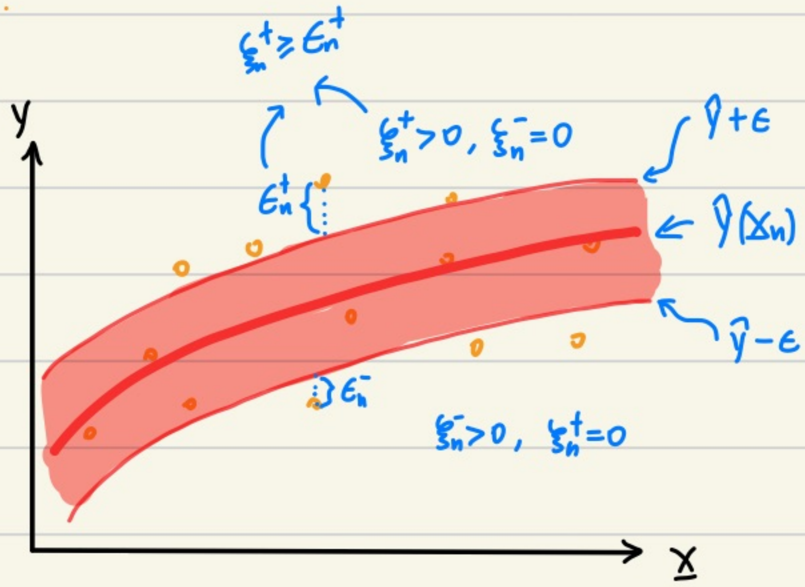

support vectors are those outside of tube

Similarly: add slack and constraints.

Inside tube: $\xi_n^+ = \xi_n^- = 0$

$\xi_n^+ > 0,\ \xi_n^- = 0$ iff $y_n > \hat{y} + \epsilon$, $\therefore y_n \leq \hat{y} + \epsilon + \xi_n^+$

$\xi_n^- > 0,\ \xi_n^+ = 0$ iff $y_n < \hat{y} - \epsilon$, $\therefore y_n \geq \hat{y} - \epsilon - \xi_n^-$




Now, $J(\underline{w}) = C\sum_{n=1}^{N}(\xi_n^+ + \xi_n^-) + \frac{1}{2}\|\underline{w}\|_2^2$

with constraints: $\begin{cases} \xi_n^+ \geq 0,\ \xi_n^- \geq 0 & \forall n \\ y_n \leq \hat{y}(x_n) + \epsilon + \xi_n^+ & \forall n \\ y_n \geq \hat{y}(x_n) - \epsilon - \xi_n^- & \forall n \end{cases}$

# Derive Lagrange Optimization Equation (Primal form)

$$L(\underline{w}, w_0, \underline{\xi}^+, \underline{\xi}^-, \underline{\mu}^+, \underline{\mu}^-, \underline{\lambda}^+, \underline{\lambda}^-) = C \cdot \sum_{n=1}^{N} (\xi_n^+ + \xi_n^-) + \frac{1}{2}\|\underline{w}\|_2^2 - \sum_{n=1}^{N} (\underbrace{\mu_n^+}_{\text{multiplier}} \xi_n^+ + \mu_n^- \xi_n^-)$$

$$- \sum_{n=1}^{N} \lambda_n^+ (\hat{y}_n + \epsilon + \xi_n^+ - y_n) - \sum_{n=1}^{N} \lambda_n^- (\epsilon + \xi_n^- - \hat{y}_n + y_n)$$

in which $\hat{y}_n = \hat{y}(\underline{x}_n) = \underline{w}^T \underline{\phi}(\underline{x}_n) + w_0$,

KKT conditions
$$\begin{cases} \xi_n^+ \geq 0,\ \xi_n^- \geq 0 \quad \forall n & \mu_n^+ \xi_n^+ = 0 \quad \forall n \\ y_n \leq \hat{y}(\underline{x}_n) + \epsilon + \xi_n^+ \quad \forall n & \mu_n^- \xi_n^- = 0 \quad \forall n \\ y_n \geq \hat{y}(\underline{x}_n) - \epsilon - \xi_n^- \quad \forall n & \lambda_n^+ (\hat{y}_n + \epsilon + \xi_n^+ - y_n) = 0 \quad \forall n \\ \mu_n^+, \mu_n^-, \lambda_n^+, \lambda_n^- \geq 0 \quad \forall n & \lambda_n^- (\epsilon + \xi_n^- - \hat{y}_n + y_n) = 0 \quad \forall n \end{cases}$$

# Derive Dual Representation by $\min_{\underline{w}, w_0, \underline{\xi}^+, \underline{\xi}^-} L(\underline{w}, w_0, \underline{\xi}^+, \underline{\xi}^-, \underline{\mu}^+, \underline{\mu}^-, \underline{\lambda}^+, \underline{\lambda}^-)$

$$\nabla_{\underline{w}} L = \underline{w} - \sum_{n=1}^{N} \lambda_n^+ \phi(\underline{x}_n) + \sum_{n=1}^{N} \lambda_n^- \phi(\underline{x}_n) = 0 \implies \underline{w} = \sum_{n=1}^{N} (\lambda_n^+ - \lambda_n^-) \underline{\phi}(\underline{x}_n)$$

$$\nabla_{w_0} L = -\sum_{n=1}^{N} \lambda_n^+ + \sum_{n=1}^{N} \lambda_n^- = 0 \implies \sum_{n=1}^{N} (\lambda_n^+ - \lambda_n^-) = 0$$

$$\nabla_{\xi_n^+} L = C - \mu_n^+ - \lambda_n^+ = 0 \implies \mu_n^+ + \lambda_n^+ = C$$

$$\nabla_{\xi_n^-} L = C - \mu_n^- - \lambda_n^- = 0 \implies \mu_n^- + \lambda_n^- = C$$

$$L = C \cdot \sum_{n=1}^{N} (\xi_n^+ + \xi_n^-) + \frac{1}{2}\|\underline{w}\|_2^2 - \sum_{n=1}^{N} (\mu_n^+ \xi_n^+ + \mu_n^- \xi_n^-) - \sum_{n=1}^{N} \lambda_n^+ (\hat{y}_n + \epsilon + \xi_n^+ - y_n) - \sum_{n=1}^{N} \lambda_n^- (\epsilon + \xi_n^- - \hat{y}_n + y_n)$$

$$= \cancel{C \cdot \sum_{n=1}^{N} (\xi_n^+ + \xi_n^-)} + \frac{1}{2} \sum_{n=1}^{N} (\lambda_n^+ - \lambda_n^-) \underline{\phi}^T(\underline{x}_n) \cdot \sum_{n=1}^{N} (\lambda_n^+ - \lambda_n^-) \underline{\phi}(\underline{x}_n) \cancel{- \sum_{n=1}^{N} \underbrace{(\mu_n^+ + \lambda_n^+)}_{C} \xi_n^+ + \underbrace{(\mu_n^- + \lambda_n^-)}_{C} \xi_n^-}$$

$$- \sum_{n=1}^{N} \lambda_n^+ \hat{y}_n - \lambda_n^- \hat{y}_n + \epsilon(\lambda_n^+ + \lambda_n^-) - y_n(\lambda_n^+ - \lambda_n^-)$$

$$= -\epsilon \sum_{n=1}^{N} (\lambda_n^+ + \lambda_n^-) + \sum_{n=1}^{N} y_n (\lambda_n^+ - \lambda_n^-) + \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} (\lambda_n^+ - \lambda_n^-)(\lambda_m^+ - \lambda_m^-) \underline{\phi}^T(\underline{x}_n) \underline{\phi}(\underline{x}_m)$$

$$\underbrace{- \sum_{n=1}^{N} (\lambda_n^+ - \lambda_n^-) \left\{ \left[ \sum_{n=1}^{N} (\lambda_n^+ - \lambda_n^-) \underline{\phi}(\underline{x}_n) \right]^T \underline{\phi}(\underline{x}_n) + w_0 \right\}}$$

$$= - \sum_{n=1}^{N} \sum_{m=1}^{N} (\lambda_n^+ - \lambda_n^-)(\lambda_m^+ - \lambda_m^-) \underline{\phi}(\underline{x}_n)^T \underline{\phi}(\underline{x}_m) - w_0 \underbrace{\sum_{n=1}^{N} (\lambda_n^+ - \lambda_n^-)}_{=0}$$

$$\boxed{\therefore L_D(\underline{\lambda}^+, \underline{\lambda}^-) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(\lambda_n^+-\lambda_n^-)(\lambda_m^+-\lambda_m^-)\underbrace{\underline{\phi}(\underline{x}_n)^T\underline{\phi}(\underline{x}_m)}_{K(\underline{x}_n,\,\underline{x}_m)\ \leftarrow\ \text{kernel}} - \epsilon\sum_{n=1}^{N}(\lambda_n^++\lambda_n^-) + \sum_{n=1}^{N} y_n(\lambda_n^+-\lambda_n^-)}$$

$\underline{\mu}^+$ & $\underline{\mu}^-$ disappear, but constraints exist

because $\mu_n^+ + \lambda_n^+ = C$ & $\mu_n^- + \lambda_n^- = C$

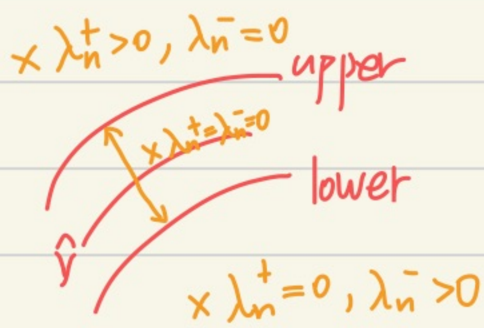
$\Rightarrow 0 \le \lambda_n^+ \le C$ & $0 \le \lambda_n^- \le C$ ← similar to SVC case.

Now KKT conditions
$$\left\{\begin{array}{llll}
\xi_n^+ \ge 0,\ \xi_n^- \ge 0 & \forall n & (C-\lambda_n^+)\xi_n^+ = 0 & \forall n \\
y_n \le \hat{y}(\underline{x}_n) + \epsilon + \xi_n^+ & \forall n & (C-\lambda_n^-)\xi_n^- = 0 & \forall n \\
y_n \ge \hat{y}(\underline{x}_n) - \epsilon - \xi_n^- & \forall n & \lambda_n^+(\hat{y}_n + \epsilon + \xi_n^+ - y_n) = 0 & \forall n \\
(C-\lambda_n^+), (C-\lambda_n^-), \lambda_n^+, \lambda_n^- \ge 0 & \forall n & \lambda_n^-(\epsilon + \xi_n^- - \hat{y}_n + y_n) = 0 & \forall n
\end{array}\right.$$

$$\sum_{n=1}^{N}(\lambda_n^+ - \lambda_n^-) = 0$$

if $\underline{x}$ is on/above upper bound, $\lambda_n^+ > 0$, else $= 0$
if $\underline{x}$ is on/below lower bound, $\lambda_n^- > 0$, else $= 0$

$\therefore$ 3 cases




SMO : $\max_{\underline{\lambda}^+, \underline{\lambda}^-} L_D(\underline{\lambda}^+, \underline{\lambda}^-)$

$$L_D(\underline{\lambda}^+, \underline{\lambda}^-) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(\lambda_n^+-\lambda_n^-)(\lambda_m^+-\lambda_m^-)K(\underline{x}_n, \underline{x}_m) - \epsilon\sum_{n=1}^{N}(\lambda_n^++\lambda_n^-) + \sum_{n=1}^{N} y_n(\lambda_n^+-\lambda_n^-)$$

Similar to SVC, choose $\lambda_m^{+/-}, \lambda_n^{+/-}$ corresponding to two points.

then
$$L_D(\lambda_m^{+/-}, \lambda_n^{+/-}) = -\frac{1}{2}\underbrace{(\lambda_m^+-\lambda_m^-)(\lambda_m^+-\lambda_m^-)K_{mm}}_{m,m} - \frac{1}{2}\underbrace{(\lambda_m^+-\lambda_m^-)(\lambda_n^+-\lambda_n^-)K_{mn}}_{m,n}$$

$$-\frac{1}{2}\underbrace{(\lambda_n^+-\lambda_n^-)(\lambda_m^+-\lambda_m^-)K_{nm}}_{n,\,m} - \frac{1}{2}\underbrace{(\lambda_n^+-\lambda_n^-)(\lambda_n^+-\lambda_n^-)K_{nn}}_{n,\,n} - \frac{1}{2}\underbrace{(\lambda_m^+-\lambda_m^-)\sum_{j=1}^{N}(\lambda_j^+-\lambda_j^-)K_{mj}}_{m,\ j\neq m,n}$$

$$-\frac{1}{2}\underbrace{(\lambda_n^+-\lambda_n^-)\sum_{j=1}^{N}(\lambda_j^+-\lambda_j^-)K_{nj}}_{n,\ j\neq m,n} - \frac{1}{2}\underbrace{(\lambda_m^+-\lambda_m^-)\sum_{i=1}^{N}(\lambda_i^+-\lambda_i^-)K_{im}}_{i\neq m,n,\ m} - \frac{1}{2}\underbrace{(\lambda_n^+-\lambda_n^-)\sum_{i=1}^{N}(\lambda_i^+-\lambda_i^-)K_{in}}_{i\neq m,n,\ n}$$

$$-\epsilon(\lambda_m^++\lambda_m^-+\lambda_n^++\lambda_n^-) + y_m(\lambda_m^+-\lambda_m^-) + y_n(\lambda_n^+-\lambda_n^-) + \underbrace{K}_{\text{constant}}$$

$$\begin{aligned} L_D(\lambda_m^{+/-},\lambda_n^{+/-}) = &-\frac{1}{2}(\lambda_m^+-\lambda_m^-)^2K_{mm} - \frac{1}{2}(\lambda_n^+-\lambda_n^-)^2K_{nn} - (\lambda_m^+-\lambda_m^-)(\lambda_n^+-\lambda_n^-)K_{mn} \\ &-(\lambda_m^+-\lambda_m^-)\underbrace{\sum_{i=1}^{N}(\lambda_i^+-\lambda_i^-)K_{im}}_{(i\neq m,n)} - (\lambda_n^+-\lambda_n^-)\underbrace{\sum_{i=1}^{N}(\lambda_i^+-\lambda_i^-)K_{in}}_{(i\neq m,n)} \\ &-\epsilon(\lambda_m^++\lambda_m^-+\lambda_n^++\lambda_n^-) + y_m(\lambda_m^+-\lambda_m^-) + y_n(\lambda_n^+-\lambda_n^-) + \underbrace{K}_{\text{constant}} \end{aligned}$$

Form dual representation KKT: $\sum_{n=1}^{N}(\lambda_n^+-\lambda_n^-)=0$

$$\therefore\ (\lambda_m^+-\lambda_m^-)+(\lambda_n^+-\lambda_n^-) = -\underbrace{\sum_{i=1}^{N}(\lambda_i^+-\lambda_i^-)}_{(i\neq m,n)} \leftarrow \text{let it be } \alpha$$

$$\therefore\ (\lambda_m^+-\lambda_m^-) = \alpha-(\lambda_n^+-\lambda_n^-)$$

$\lambda_m^++\lambda_m^- = |\alpha-(\lambda_n^+-\lambda_n^-)|$ because only $\lambda_m^+$ or $\lambda_m^- > 0$

Also, let $V_m = \underbrace{\sum_{i=1}^{N}(\lambda_i^+-\lambda_i^-)K_{im}}_{(i\neq m,n)}$ $\quad V_n = \underbrace{\sum_{i=1}^{N}(\lambda_i^+-\lambda_i^-)K_{in}}_{(i\neq m,n)}$

$$\begin{aligned} L_D(\lambda_n^+,\lambda_n^-) = &-\frac{1}{2}(\alpha-\lambda_n^++\lambda_n^-)^2K_{mm} - \frac{1}{2}(\lambda_n^+-\lambda_n^-)^2K_{nn} - (\alpha-\lambda_n^++\lambda_n^-)(\lambda_n^+-\lambda_n^-)K_{mn} \\ &-(\alpha-\lambda_n^++\lambda_n^-)V_m - (\lambda_n^+-\lambda_n^-)V_n - \epsilon\cdot|\alpha-(\lambda_n^+-\lambda_n^-)| - \epsilon(\lambda_n^++\lambda_n^-) \\ &+y_m(\alpha-\lambda_n^++\lambda_n^-) + y_n(\lambda_n^+-\lambda_n^-) + K \end{aligned}$$

Combine two variable $\lambda_i^+$ & $\lambda_i^-$ into one $\beta_i$ $\begin{cases} \beta_i = \lambda_i^+-\lambda_i^- \\ |\beta_i| = \lambda_i^++\lambda_i^- \end{cases}$

Then, $L_D(\beta_n) = -\frac{1}{2}(\alpha-\beta_n)^2 K_{mm} - \frac{1}{2}\beta_n^2 K_{nn} - (\alpha-\beta_n)\beta_n K_{mn}$

$- (\alpha-\beta_n)V_m - \beta_n V_n - \epsilon\cdot|\alpha-\beta_n| - \epsilon\cdot|\beta_n| + y_m(\alpha-\beta_n) + y_n\beta_n + K$

$$\frac{\partial L_D(\beta_n)}{\partial \beta_n} = (\alpha-\beta_n)K_{mm} - \beta_n K_{nn} + \beta_n K_{mn} - (\alpha-\beta_n)K_{mn} + V_m - V_n$$
$$+ \epsilon\cdot \text{sign}(\beta_m) - \epsilon\cdot\text{sign}(\beta_n) - y_m + y_n$$

where $\text{sign}(x) = \begin{cases} 1, & x>0 \\ 0, & x=0 \\ -1, & x<0 \end{cases}$

Let $\frac{\partial L_D(\beta_n)}{\partial \beta_n} = 0 \Rightarrow \beta_n K_{mm} + \beta_n K_{nn} - 2\beta_n K_{mn} = \alpha K_{mm} - \alpha K_{mn} - y_m + y_n$
$+ \epsilon[\text{sign}(\beta_m) - \text{sign}(\beta_n)] + V_m - V_n$

because $V_m = \sum_{\substack{i=1 \\ (i\neq m,n)}}^{N}(\lambda_i^+ - \lambda_i^-)K_{im}$ $\quad V_n = \sum_{\substack{i=1 \\ (i\neq m,n)}}^{N}(\lambda_i^+ - \lambda_i^-)K_{in}$ $\quad$ & $\underline{w} = \sum_{n=1}^{N}(\lambda_n^+ - \lambda_n^-)\underline{\phi}(\underline{x}_n)$

& $f(\underline{x}_k) = \underline{w}^T\underline{x}_k + w_0$
$= \sum_{n=1}^{N}(\lambda_n^+ - \lambda_n^-)K_{nk} + w_0$

$\therefore V_i = f(\underline{x}_i) - \beta_m K_{mi} - \beta_n K_{ni} - w_0$

$\therefore V_m - V_n = f(\underline{x}_m) - f(\underline{x}_n) - \beta_m(K_{mm} - K_{mn}) - \beta_n(K_{nm} - K_{nn})$

$$\therefore \underbrace{\beta_n}_{\text{new}}(K_{mm} + K_{nn} - 2K_{mn}) = \alpha(K_{mm} - K_{mn}) - y_m + y_n + \epsilon[\text{sign}(\beta_m) - \text{sign}(\beta_n)]$$
$$+ f(\underline{x}_m) - f(\underline{x}_n) - \underbrace{\beta_m}_{\text{old}}(K_{mm} - K_{mn}) - \underbrace{\beta_n}_{\text{old}}(K_{nm} - K_{nn})$$

because $\alpha = -\sum_{\substack{i=1 \\ (i\neq m,n)}}^{N}(\lambda_i^+ - \lambda_i^-) = \underbrace{\beta_m + \beta_n}_{\text{old}}$

$$\therefore \beta_n^{new}\cdot\eta = y_n - y_m + \epsilon[\text{sign}(\beta_m) - \text{sign}(\beta_n)] + f(\underline{x}_m) - f(\underline{x}_n) + \beta_n^{old}\cdot\eta$$

in which $\eta = (K_{mm} + K_{nn} - 2K_{mn})$

$$\Rightarrow \boxed{\begin{array}{l} \beta_n^{new} = \beta_n^{old} + \frac{1}{\eta}\left(y_n - y_m + \epsilon\left[\mathrm{sign}(\beta_m) - \mathrm{sign}(\beta_n)\right] + f(x_m) - f(x_n)\right) \\ \text{where } \begin{cases} f(x_k) = \sum_{n=1}^{N}(\lambda_n^+ - \lambda_n^-)K_{nk} + w_0 = \sum_{n=1}^{N}\beta_n^{old}K_{nk} + w_0 \\ \eta = (K_{mm} + K_{nn} - 2K_{mn}) \end{cases} \end{array}}$$

Note*: In above $\beta_n$ update function, the $\beta_m$ & $\beta_n$ in sign(·) function should be new instead of old. But this makes it a recursive function.

Finding solutions ☆ [Efficient SVM Regression Training with SMO. Gary & Steve. 2002]

① if $\beta_m < 0 < \beta_n$, $\mathrm{sign}(\beta_m) - \mathrm{sign}(\beta_n) = -2$

② if $\beta_n < 0 < \beta_m$, $\mathrm{sign}(\beta_m) - \mathrm{sign}(\beta_n) = 2$

③ if $\beta_m, \beta_n \overset{\text{both}}{\gtrless} 0$, $\mathrm{sign}(\beta_m) - \mathrm{sign}(\beta_n) = 0$

In these two cases, we don't really do this

if $\beta_m = 0$, $\beta_n = \alpha - \beta_m = \alpha = \beta_m^{old} + \beta_n^{old}$

$\mathrm{sign}(\beta_m) - \mathrm{sign}(\beta_n) = -\mathrm{sign}(\beta_n) = -\mathrm{sign}(\alpha)$

if $\beta_n = 0$, $\beta_m = \alpha - \beta_n = \alpha = \beta_m^{old} + \beta_n^{old}$

$\mathrm{sign}(\beta_m) - \mathrm{sign}(\beta_n) = \mathrm{sign}(\beta_m) = \mathrm{sign}(\alpha)$

Because sign(·) function makes the update function a recursive function, there is no way to calculate; below is a reasonable efficient algorithm that is proved by exps.

For a single step of SMO (update $\beta$)

1 we have $\alpha = \beta_m^{old} + \beta_n^{old}$

$\eta = K_{mm} + K_{nn} - 2K_{mn}$

$\Delta = \frac{2\epsilon}{\eta}$ ← correction term [① ②]

2 $\beta_n = \beta_n^{old} + \frac{1}{\eta}\left[y_n - y_m + f(\underline{x}_m) - f(\underline{x}_n)\right]$ with old $\beta$

$\beta_m = \alpha - \beta_n$ ← if sign($\beta_m$) == sign($\beta_n$), jump 4. [③]

3 if $\beta_m \cdot \beta_n < 0$, use correction term.

if $(|\beta_m| \geq \Delta$ && $|\beta_n| \geq \Delta)$ ← sign won't be affected by $\Delta$ term

$\beta_n = \beta_n + \text{sign}(\beta_m) \cdot \Delta$

else ← signs of $\beta_m$ & $\beta_n$ will be affected, one of them is set to be 0, another be $\alpha$.

$\beta_n = \alpha$ if $|\beta_n| > |\beta_m|$ else 0

4 Crop step by $[0 \leq \lambda_i^+, \lambda_i^- \leq C]$ $[-C \leq \beta_i = \lambda_i^+ - \lambda_i^- \leq C]$

$$L = \max(\alpha - C, -C)$$

$$H = \min(C, \alpha + C)$$

$$\beta_n^{new} = \min(\max(\beta_n, L), H)$$

$$\beta_m^{new} = \alpha - \beta_n^{new}$$

## 5 update $w_0$, to make it easy:

if $\beta_m^{new} = 0$ (within tube), force it have $f(\underline{x}_m) = y_m$, then $y_m = \sum_{i=1}^{N} \beta_i^{new} K_{im} + w_0^{new}$

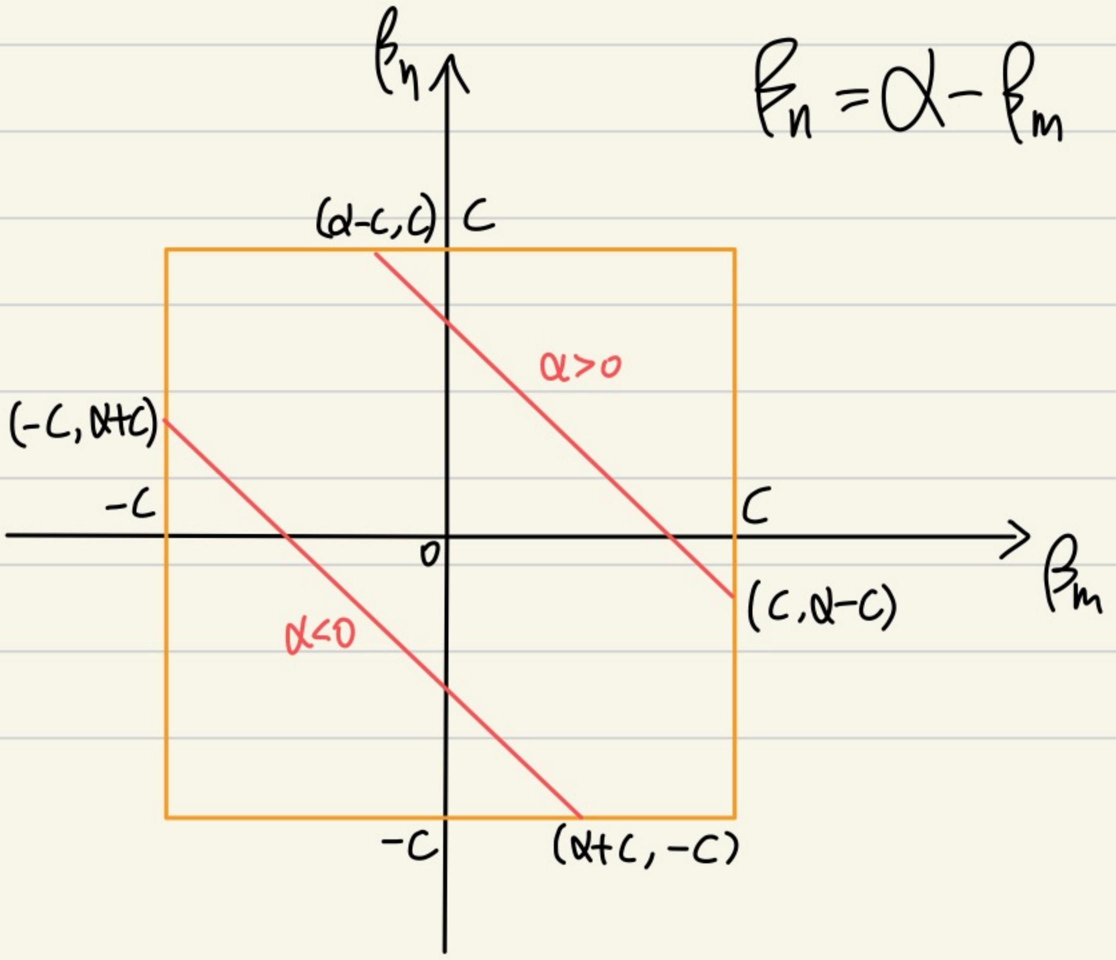
i) $$w_0^{new} = y_m - \sum_{i=1}^{N}{}^{(i \neq m,n)} \beta_i^{new} K_{im} - \beta_m^{new} K_{mm} - \beta_n^{new} K_{nm}$$
$$= y_m - [f^{old}(\underline{x}_m) - w_0^{old} - \beta_m^{old} K_{mm} - \beta_n^{old} K_{nm}] - \beta_m^{new} K_{mm} - \beta_n^{new} K_{nm}$$
$$= w_0^{old} + y_m - f^{old}(\underline{x}_m) + (\beta_m^{old} - \beta_m^{new}) K_{mm} + (\beta_n^{old} - \beta_n^{new}) K_{mn}$$

if $\beta_n^{new} = 0$ (within tube), force it have $f(\underline{x}_n) = y_n$

ii) $$w_0^{new} = w_0^{old} + y_n - f^{old}(\underline{x}_n) + (\beta_m^{old} - \beta_m^{new}) K_{mn} + (\beta_n^{old} - \beta_n^{new}) K_{nn}$$

if neither, average over i) & ii).

## Crop Visualization




$\therefore$ $\alpha < 0$, $-C \leq \beta_n \leq \alpha + C$

$\alpha > 0$, $\alpha - C \leq \beta_n \leq C$

$\Rightarrow L = \max(\alpha - C, -C)$

$H = \min(\alpha + C, C)$

for $\beta_n$