Support Vector Regression

JrPhy

Introduction

- Now we’ve known the regularization is

$$\min_{w}\left(\frac{\lambda}{N}ww^{T}+\frac{1}{N}\sum_{i}err_{i}\right)$$

- And by the representation theorem, it can be represent as

$$\min_{w}\left(\frac{\lambda}{N}ww^{T}+\frac{1}{N}\sum_{i}err_{i}\right)=\min_{w}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\beta_{i}\beta_{j}K(x_{i},x_{j})+\frac{1}{N}\sum_{i=1}^{N}\log\left(1+\exp\left(-y_{n}\sum_{j=1}^{N}\beta_{j}K(x_{i},x_{j})\right)\right)\right)$$

- Let’s apply square error in the err_i and derive the analytical solution.

Optimize

- The square error is $(y_n - w^T z_n)^2$, so we want to minimize

$$\min_w \left(\frac{\lambda}{N} w_{opt} w_{opt}{}^T + \frac{1}{N} \sum_i \left(y_n - w_{opt}{}^T z_n \right)^2 \right)$$

- Where $w_{opt} = \sum_{i=1}^{N} \beta_i z_i$, and $z_n = \Phi(x_n)$, so

$$\frac{\lambda}{N} w_{opt} w_{opt}{}^T + \frac{1}{N} \sum_i \left(y_n - w_{opt}{}^T z_n \right)^2$$

$$= \frac{\lambda}{N} \sum_{i=1}^{n} \sum_{j=1}^{m} \beta_i \beta_j K(x_i, x_j) + \frac{1}{N} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} \beta_i K(x_i, x_j) \right)^2$$

Optimize

$$\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{m}\beta_i K(x_i,x_j)\right)^2=\left\|Y-\beta K\right\|^2=YY^T+\beta KK^T\beta^T-2\beta^T KY$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_i\beta_j K(x_i,x_j)=\beta K\beta^T$$

- So what we want to minimize is

$$E(\beta)=\frac{\lambda}{N}\beta K\beta^T+\frac{1}{N}\left(YY^T+\beta KK^T\beta^T-2\beta^T KY\right)$$

$$\nabla E(\beta)=\frac{2\lambda}{N}\beta K+\frac{1}{N}\left(2\beta KK^T-2KY\right)=\frac{2}{N}K\left(\lambda\beta+\beta K^T-Y\right)$$

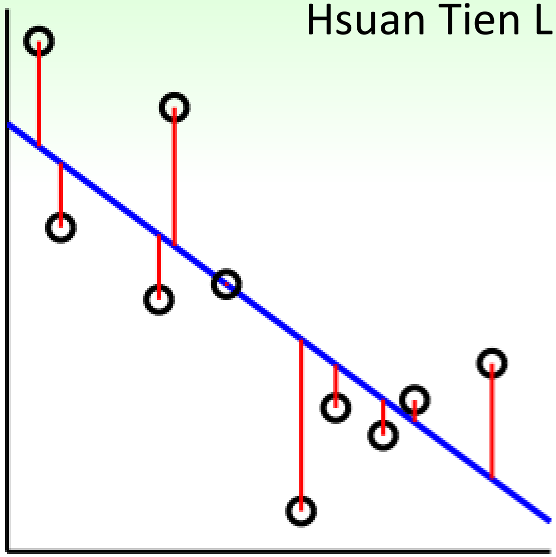
$$\boxed{=\frac{2}{N}K\left(\beta\left(\lambda I+K^T\right)-Y\right)}$$

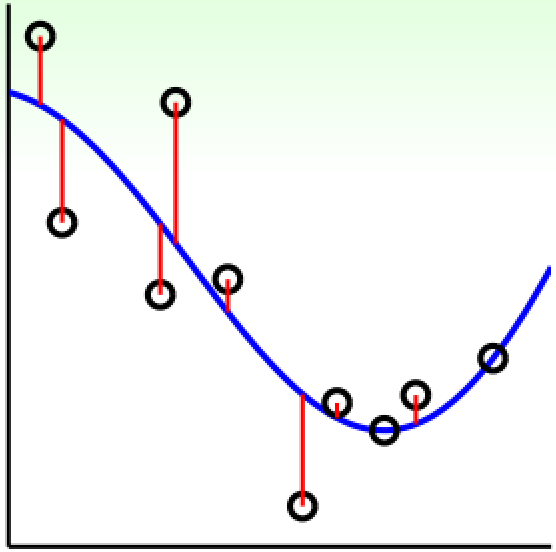
least-square SVM

- So the analytical solution is
    - $\beta = (\lambda I+K)^{-1}Y$
- By the Mercer's condition, K is semi-definite, for all $\lambda > 0$, the inverse of $\lambda I+K$ exists.
- So far we can not only do the linear regression, but the kernel regression for the SVM
- If the error is the square error, then the SVM is called **least-square SVM(LSSVM)**, or called **kernel ridge regression**.

Linear versus Kernel Ridge Regression

Hsuan Tien Lin, mltech/206_handout





linear ridge regression

$$\mathbf{w} = (\lambda \mathrm{I} + \mathrm{X}^T \mathrm{X})^{-1} \mathrm{X}^T \mathbf{y}$$

- more restricted
- $O(d^3 + d^2 N)$ training; $O(d)$ prediction —**efficient when $N \gg d$**

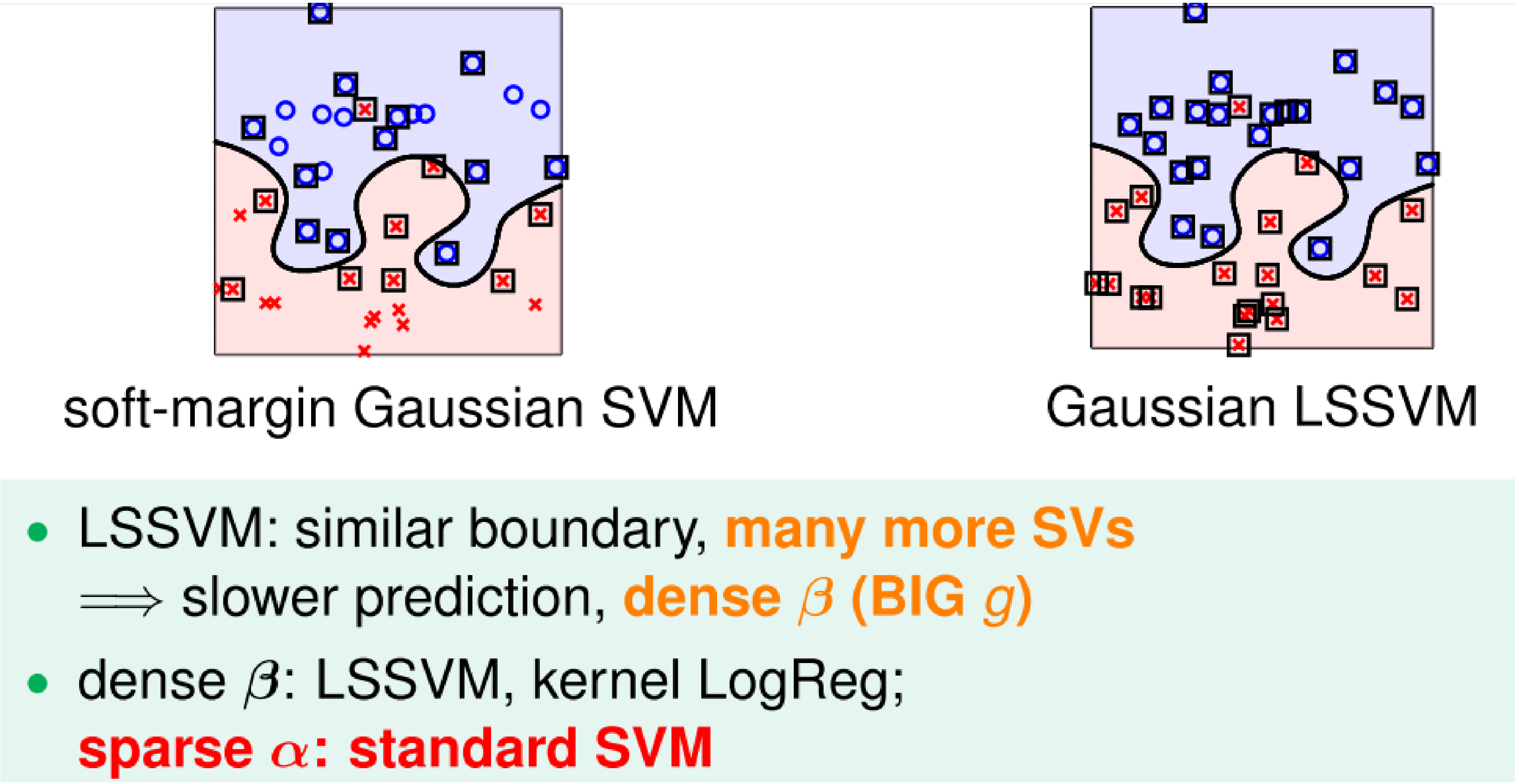
kernel ridge regression

$$\beta = (\lambda \mathrm{I} + \mathrm{K})^{-1} \mathbf{y}$$

- **more flexible** with K
- $O(N^3)$ training; $O(N)$ prediction —hard for big data

Combine regression and soft-margin

- The soft-margin allows some mistakes in the region, so that there is less SV than the hard margin, and it’s more efficiency, so what we want is to combine the regression and soft-margin.



soft-margin Gaussian SVM

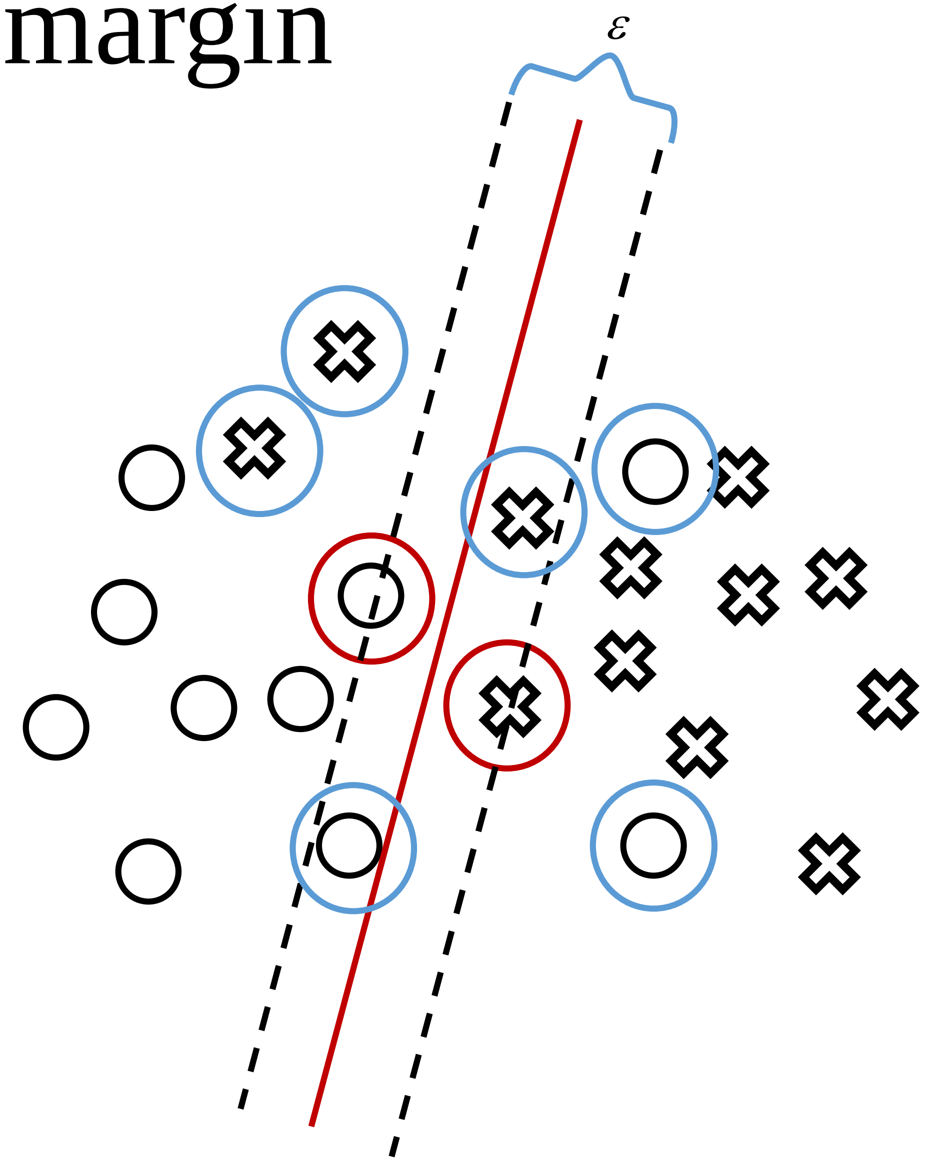
Gaussian LSSVM

- LSSVM: similar boundary, **many more SVs**
  $\Longrightarrow$ slower prediction, **dense β (BIG g)**
- dense β: LSSVM, kernel LogReg;
  sparse α: standard SVM

Hsuan Tien Lin, mltech/206_handout

Combine regression and soft-margin

- Suppose the red line is determined by the regression, and the area between the dash line is the boundary and its width is ε . We don’t care about the error in the area.
- So the $err(y, s) = \max(0, |y – s| – \varepsilon)$
- $|y – s| – \varepsilon < 0$: $err(y, s) = 0$
- $|y – s| – \varepsilon > 0$: $err(y, s) = |y – s| – \varepsilon$
- Call it tube regression.



Combine regression and soft-margin

- So the error is linear, it grows slower than the square error, and it’s close to the square error when the *s* is small.
- Now we want to minimize the

$$\min_{w}\left(\frac{\lambda}{N}ww^{T}+\frac{1}{N}\sum_{i=1}^{n}\max(0,\left|wz_{i}-y\right|-\varepsilon)\right)$$

- But it can’t differential at some points, so let’s mimic the standard SVM problem

$$\min_{b,w}\left(\frac{1}{2}ww^{T}+C\sum_{i=1}^{n}\max(0,\left|wz_{i}+b-y\right|-\varepsilon)\right)$$

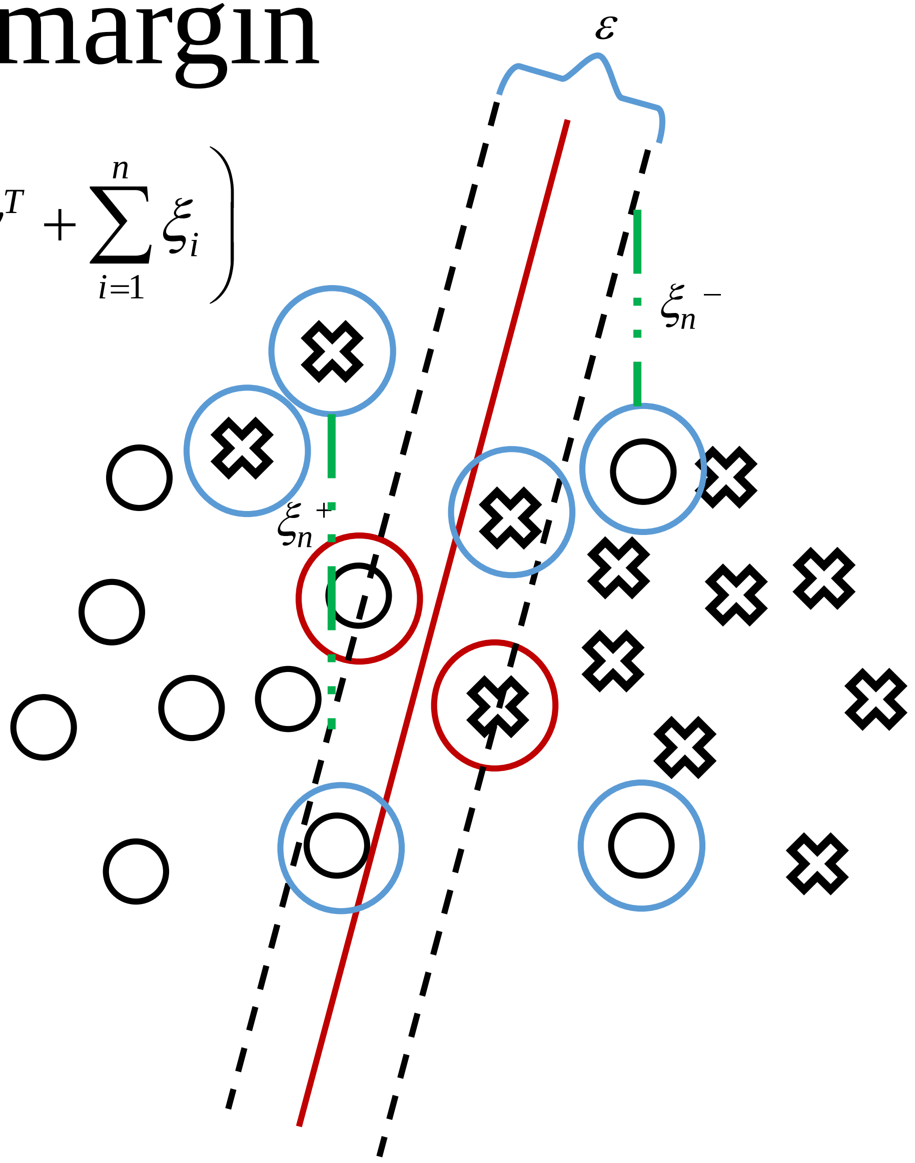
- The symbols are the same as the derivation in SVM.

Combine regression and soft-margin

$$\min_{b,w}\left(\frac{1}{2}ww^T + C\sum_{i=1}^{n}\max(0,|wz_i+b-y|-\varepsilon)\right) \rightarrow \min_{b,w,\xi}\left(\frac{1}{2}ww^T+\sum_{i=1}^{n}\xi_i\right)$$

- constrains: $|\, y - wz_n + b| \le \varepsilon + \xi_n$, $\xi_n \ge 0$
- But there is a absolute value, it’s not a QP problem, so we separate it as
- $-\varepsilon - \xi_n^{-} \le y - (wz_n + b) \le \varepsilon + \xi_n^{+}$,
- $\xi_n^{+} \ge 0$, $\xi_n^{-} \ge 0$
- Then what we want to minimize is

$$\min_{b,w,\xi}\left(\frac{1}{2}ww^T+\sum_{i=1}^{n}\xi_i\right) \rightarrow \min_{b,w,\xi^-,\xi^+}\left(\frac{1}{2}ww^T+\sum_{i=1}^{n}\left(\xi_i^{-}+\xi_i^{+}\right)\right)$$



Solving the Lagrange undetermined multiplier

- The Lagrange function is

$$L(\alpha^+, \alpha^-, w, b, \xi^+, \xi^-, \beta^+, \beta^-) = \frac{1}{2}\|w\|^2 + C\sum_i \left(\xi_i^+ + \xi_i^-\right) - \sum_i \left(\beta_i^+ \xi_i^+ + \beta_i^- \xi_i^-\right)$$

$$+\sum_i \alpha_i^+ [\left(y_i - w^T z_i - b\right) - \left(\varepsilon + \xi_i^+\right)] + \sum_i \alpha_i^- [-\left(\varepsilon + \xi_i^-\right) - \left(y_i - w^T z_i - b\right)]$$

$$\frac{\partial}{\partial w_i} L = w - \sum_{i=1}^{n} \alpha_i^+ z_i + \sum_{i=1}^{n} \alpha_i^- z_i = 0 \rightarrow w = \sum_{i=1}^{n} \left(\alpha_i^+ - \alpha_i^-\right) z_i$$

$$\frac{\partial}{\partial \xi_i^+} L = 0 = C - \beta_i^+ - \alpha_i^+ \qquad \frac{\partial}{\partial \xi_i^-} L = 0 = C - \beta_i^- - \alpha_i^-$$

$$\frac{\partial}{\partial b} L = 0 \rightarrow \sum_{i=1}^{n} \left(\alpha_i^+ - \alpha_i^-\right) = 0$$

Solving the Lagrange undetermined multiplier

$$L(\alpha^+,\alpha^-,w,b,\xi^+,\xi^-,\beta^+,\beta^-)=\frac{1}{2}\|w\|^2+C\sum_i\left(\xi_i^+ +\xi_i^-\right)-\sum_i\left(\beta_i^+\xi_i^+ +\beta_i^-\xi_i^-\right)$$

$$+\sum_i\alpha_i^+[\left(y_i-w^Tz_i-b\right)-\left(\varepsilon+\xi_i^+\right)]+\sum_i\alpha_i^-[-\left(\varepsilon+\xi_i^-\right)-\left(y_i-w^Tz_i-b\right)]$$

- Substitute all condition, then we get

$$L(\alpha^+,\alpha^-,w,b,\xi^+,\xi^-,\beta^+,\beta^-)$$

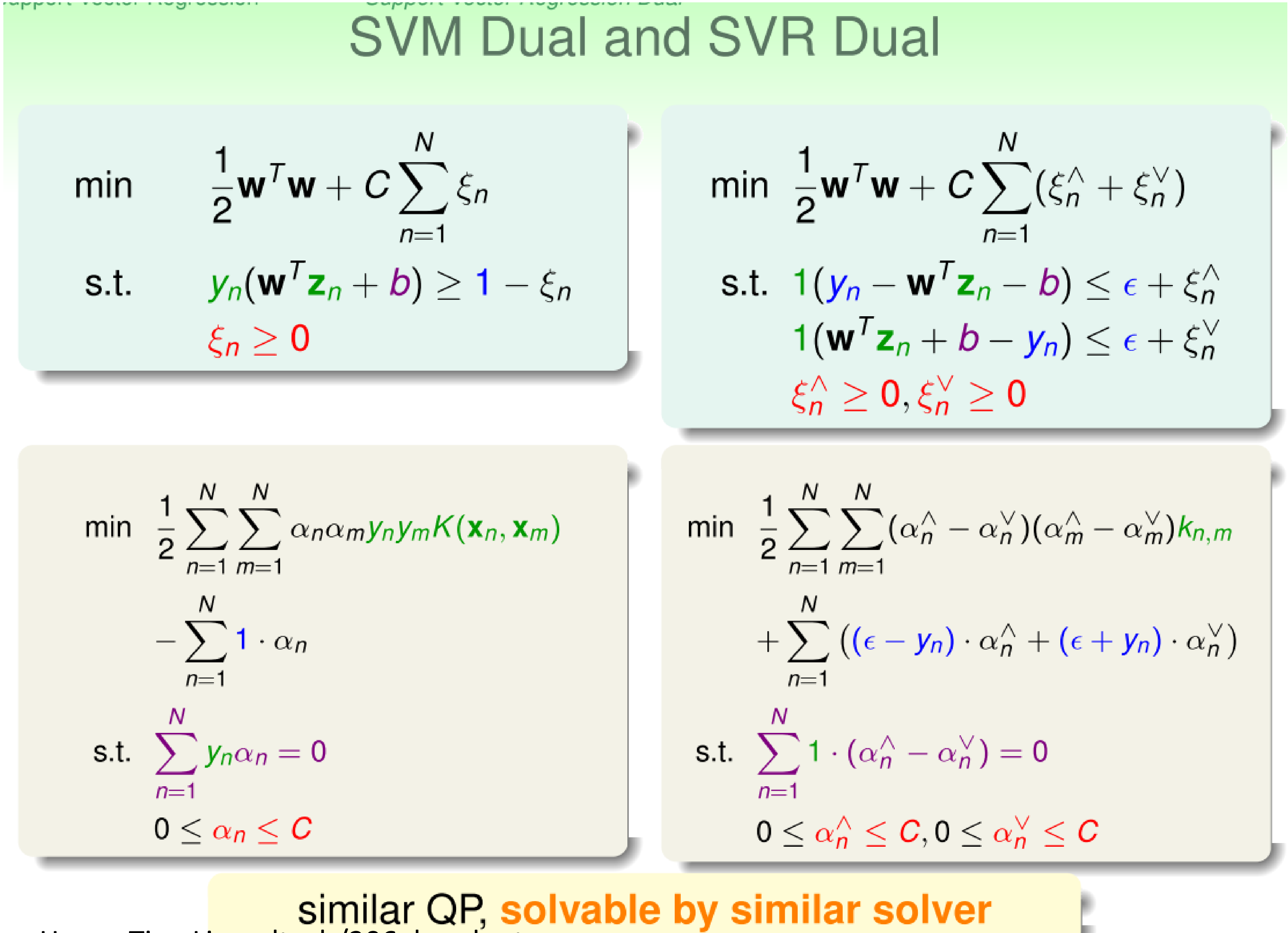
$$=\sum_{i=1}^n\xi_i^+\left(C-\alpha_i^+-\beta_i^+\right)+\sum_{i=1}^n\xi_i^-\left(C-\alpha_i^--\beta_i^-\right)-\varepsilon\sum_{i=1}^n\left(\alpha_i^+-\alpha_i^-\right)+\sum_{i=1}^n\left(\alpha_i^+-\alpha_i^-\right)\left(y_i-w^Tz_i-b\right)$$

$$=\sum_{i=1}^n\left(\alpha_i^+-\alpha_i^-\right)\left(y_i-w^Tz_i-b\right)-\varepsilon\sum_{i=1}^n\left(\alpha_i^+-\alpha_i^-\right)$$

Solving the Lagrange undetermined multiplier

- By the KKT condition
- $\alpha_i^{+}[\left(y_i - w^T z_i - b\right) - \left(\varepsilon + \xi_i^{+}\right)] = 0$ and $\left(C - \alpha_i^{+}\right)\xi_i^{+} = 0$

  $\alpha_i^{-}[\left(y_i - w^T z_i - b\right) - \left(\varepsilon + \xi_i^{-}\right)] = 0$ $\left(C - \alpha_i^{-}\right)\xi_i^{-} = 0$
- If $\xi_i^{+} > 0$, then $C = \alpha_i^{+}$, $\xi_i^{-} > 0$, then $C = \alpha_i^{-}$
- So $0 \le \alpha_i^{+}, \alpha_i^{-} \le C$.



SVM Dual and SVR Dual

$$\begin{aligned} \min \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}\xi_n \\ \text{s.t.} \quad & y_n(\mathbf{w}^T\mathbf{z}_n + b) \geq 1 - \xi_n \\ & \xi_n \geq 0 \end{aligned}$$

$$\begin{aligned} \min \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}(\xi_n^{\wedge} + \xi_n^{\vee}) \\ \text{s.t.} \quad & 1(y_n - \mathbf{w}^T\mathbf{z}_n - b) \leq \epsilon + \xi_n^{\wedge} \\ & 1(\mathbf{w}^T\mathbf{z}_n + b - y_n) \leq \epsilon + \xi_n^{\vee} \\ & \xi_n^{\wedge} \geq 0, \xi_n^{\vee} \geq 0 \end{aligned}$$

$$\begin{aligned} \min \quad & \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m y_n y_m K(\mathbf{x}_n, \mathbf{x}_m) \\ & -\sum_{n=1}^{N} 1 \cdot \alpha_n \\ \text{s.t.} \quad & \sum_{n=1}^{N} y_n\alpha_n = 0 \\ & 0 \leq \alpha_n \leq C \end{aligned}$$

$$\begin{aligned} \min \quad & \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(\alpha_n^{\wedge} - \alpha_n^{\vee})(\alpha_m^{\wedge} - \alpha_m^{\vee})k_{n,m} \\ & +\sum_{n=1}^{N}\left((\epsilon - y_n)\cdot\alpha_n^{\wedge} + (\epsilon + y_n)\cdot\alpha_n^{\vee}\right) \\ \text{s.t.} \quad & \sum_{n=1}^{N} 1 \cdot (\alpha_n^{\wedge} - \alpha_n^{\vee}) = 0 \\ & 0 \leq \alpha_n^{\wedge} \leq C, 0 \leq \alpha_n^{\vee} \leq C \end{aligned}$$

similar QP, **solvable by similar solver**

Hsuan Tien Lin, mltech/206_handout

Sparsity of SVR solution

- In the soft-margin SVM, we just care about the mistakes outside the area, so as SVR.
- If the classification is correct, then the most of the mistakes are in the area, so that $\xi_n{}^+ = \xi_n{}^- = 0$, but $\alpha_i{}^+[(y_i - w^T z_i - b) - \varepsilon] = 0$ and $[(y_i - w^T z_i - b) - \varepsilon] \neq 0$, so $\alpha_i{}^+ = 0$, so as $\alpha_i{}^-$.
- So that the solution of SVR is sparsity, it's like SVM.
- There is a summary of "Map of Linear/Kernel Models" in the last part of "Hsuan Tien Lin, mltech/206_handout", it's a good reference.