Introduction to Machine Learning - CS725
Instructor: Prof. Ganesh Ramakrishnan
Lecture 11 - KKT Conditions, Support Vector Regression and its Dual

KKT conditions for the Constrained (Convex) Problem

- Here, $\mathbf{w} \in \mathbb{R}^n$ and the domain is the intersection of all functions. Lagrangian is:

$$L(\mathbf{w}, \lambda, \mu) = f(\mathbf{w}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{w}) + \sum_{j=1}^{p} \mu_j h_j(\mathbf{w})$$

- KKT **necessary** conditions for all differentiable functions (i.e. f, g_i, h_j) with optimality points $\hat{\mathbf{w}}$ and $(\hat{\lambda}, \hat{\mu})$ are:
  - $\nabla f(\hat{\mathbf{w}}) + \sum_{i=1}^{m} \hat{\lambda}_i \nabla g_i(\hat{\mathbf{w}}) + \sum_{j=1}^{p} \hat{\mu}_j \nabla h_j(\hat{\mathbf{w}}) = 0$
  - $g_i(\hat{\mathbf{w}}) \leq 0; 1 \leq i \leq m$
  - $\hat{\lambda}_i \geq 0; 1 \leq i \leq m$
  - $\hat{\lambda}_i g_i(\hat{\mathbf{w}}) = 0; 1 \leq i \leq m$
  - $h_j(\hat{\mathbf{w}}) = 0; 1 \leq j \leq p$

linear

- When f and $g_i, \forall i \in [1, m]$ are convex and $h_j, \forall j \in [1, p]$ are affine, KKT conditions are also **sufficient** for optimality at $\hat{\mathbf{w}}$ and $(\hat{\lambda}, \hat{\mu})$

KKT conditions for the Constrained (Convex) Problem
Recap Application 1: Equivalence of two forms of Ridge Regression

Equivalent Forms of Ridge Regression

- Consider the formulation in which we limit the weights of the coefficients by putting a constraint on size of the L2 norm of the weight vector:

$$\text{argmin}_{\mathbf{w}}(\mathbf{\Phi w} - \mathbf{y})^T(\mathbf{\Phi w} - \mathbf{y})$$

$$\|\mathbf{w}\|_2^2 \leq \xi$$

- The objective function, namely $f(\mathbf{w}) = (\mathbf{\Phi w} - \mathbf{y})^{\mathbf{T}}(\mathbf{\Phi w} - \mathbf{y})$ is strictly convex. The constraint function, $g(\mathbf{w}) = \|\mathbf{w}\|_2^2 - \xi$, is also convex.
- For convex $g(\mathbf{w})$, the set $\{\mathbf{w}|\mathbf{g}(\mathbf{w}) \leq \mathbf{0}\}$, is also convex. (Why?)

Equivalent Forms of Ridge Regression

- To minimize the error function subject to constraint $|\mathbf{w}| \leq \xi$, we apply KKT conditions at the point of optimality $\mathbf{w}^*$

$$\nabla_{\mathbf{w}^*}(f(\mathbf{w}) + \lambda \mathbf{g}(\mathbf{w})) = \mathbf{0}$$

$\nabla_w L(w, \lambda)$

(the first KKT condition). Here, $f(\mathbf{w}) = (\Phi\mathbf{w} - \mathbf{y})^T(\Phi\mathbf{w} - \mathbf{y})$ and, $g(\mathbf{w}) = \|\mathbf{w}\|^2 - \xi$.

- Solving we get,

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

From the second KKT condition we get,

$$\|\mathbf{w}^*\|^2 \leq \xi$$

From the third KKT condition,

$$\lambda \geq 0$$

From the fourth condition

$$\lambda\|\mathbf{w}^*\|^2 = \lambda\xi$$

if $\lambda \geq 0$, $\|w^*\|^2 = \xi$

① If by setting $\lambda = 0$ $\|(\Phi^T\Phi)^{-1}\Phi^T y\|^2 \leq \xi$

Then $\lambda = 0$ & $w^* = (\Phi^T\Phi)^{-1}\Phi^T y$ is soln

else ② Increase λ s.t $\|w^*\|_2^2$

$= \|(\Phi^T\Phi + \lambda I)^{-1}\Phi^T y\| = \xi$

Equivalent Forms of Ridge Regression

- Values of **w** and λ that satisfy all these equations would yield an optimal solution. That is, if

case 1

$$\|\mathbf{w}^*\| = \|(\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}\| \leq \xi$$

then $\lambda = 0$ is the solution. Else, for some sufficiently large value, λ will be the solution to

$$\|\mathbf{w}^*\| = \|(\Phi^T\Phi + \lambda I)^{-1}\Phi^T\mathbf{y}\| = \xi$$

Bound on λ in the regularized least square solution Case 2

- Consider,

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

We multiply $(\Phi^T\Phi + \lambda I)$ on both sides and obtain,

$$\|(\Phi^T\Phi)\mathbf{w}^* + (\lambda \mathbf{I})\mathbf{w}^*\| = \|\mathbf{\Phi^T y}\|$$

Using the triangle inequality we obtain,

$$\|(\Phi^T\Phi)\mathbf{w}^*\| + (\lambda)\|\mathbf{w}^*\| \geq \|(\mathbf{\Phi^T\Phi})\mathbf{w}^* + (\lambda\mathbf{I})\mathbf{w}^*\| = \|\mathbf{\Phi^T y}\|$$

- By the Cauchy Shwarz inequality, $\|(\Phi^T\Phi)\mathbf{w}^*\| \leq \alpha\|\mathbf{w}^*\|$ for some $\alpha = \|(\Phi^T\Phi)\|$. $\|\Phi^T\Phi\|$
Substituting in the previous equation,

$\|AB\| \leq \|A\| * \|B\|$

$$(\alpha + \lambda)\|\mathbf{w}^*\| \geq \|\mathbf{\Phi^T y}\|$$

i.e.

$$\left[\lambda \geq \frac{\|\Phi^T \mathbf{y}\|}{\|\mathbf{w}^*\|} - \alpha\right]$$

Note that when $\|\mathbf{w}^*\| \to \mathbf{0}, \lambda \to \infty$. (Any intuition?) Using $\|\mathbf{w}^*\|^2 \leq \xi$ we get,

Bound on λ in the regularized least square solution

$\|(\Phi^T\Phi)\mathbf{w}^*\| \leq \alpha\|\mathbf{w}^*\|$ for some α for finite $|(\Phi^T\Phi)\mathbf{w}^*\|$. Substituting in the previous equation,

$$(\alpha + \lambda)\|\mathbf{w}^*\| \geq \|\Phi^T\mathbf{y}\|$$

i.e.

$$\lambda \geq \frac{\|\Phi^T\mathbf{y}\|}{\|\mathbf{w}^*\|} - \alpha$$

Note that when $\|\mathbf{w}^*\| \to 0, \lambda \to \infty$. (Any intuition?) Using $\|\mathbf{w}^*\|^2 \leq \xi$ we get,

$$\lambda \geq \frac{\|\Phi^T\mathbf{y}\|}{\sqrt{\xi}} - \alpha$$

This is not the exact solution of λ but the bound proves the existence of λ for some ξ and Φ.

The Resultant alternative objective function

Substituting $g(\mathbf{w}) = \|\mathbf{w}\|^2 - \xi$, in the first KKT equation considered earlier:

$$\nabla_{\mathbf{w}^*}(f(\mathbf{w}) + \lambda \cdot (\|\mathbf{w}\|^2 - \xi)) = \mathbf{0}$$

This is equivalent to solving

$$\min(\| \Phi\mathbf{w} - \mathbf{y} \|^2 + \lambda \| \mathbf{w} \|^2)$$

for the same choice of λ. This form of **regularized** ridge regression is the **penalized ridge regression**.

KKT conditions for the Constrained (Convex) Problem Application 2: SVR and its Dual

KKT and Dual for SVR

- $$\min_{\mathbf{w},b,\xi_i,\xi_i^*} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*)$$
  s.t. $\forall i$,
  $y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b \leq \epsilon + \xi_i$, → α_i
  $b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i \leq \epsilon + \xi_i^*$, → α_i^*
  μ_i, μ_i^* ← $\xi_i, \xi_i^* \geq 0$
- Let's consider the lagrange multipliers α_i, α_i^*, μ_i and μ_i^* corresponding to the above-mentioned constraints.
- The Lagrange Function is

$$L(\underbrace{w, b, \xi_i, \xi_i^*}_{\text{Primal/original vars}}, \underbrace{\alpha_i, \alpha_i^*, \mu_i, \mu_i^*}_{\text{New vars}}) = \frac{1}{2}\|w\|^2 + C\sum_i(\xi_i + \xi_i^*)$$
$$+ \sum_i \alpha_i(y_i - w^T\phi(x_i) - b - \epsilon - \xi_i) + \sum_i \alpha_i^*(b + w^T\phi(x_i) - y_i - \epsilon - \xi_i^*)$$
$$- \sum \mu_i\xi_i - \sum \mu_i^*\xi_i^*$$

KKT and Dual for SVR

- $$\min_{\mathbf{w},b,\xi_i,\xi_i^*} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*)$$
  s.t. $\forall i$,
  $$y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b \leq \epsilon + \xi_i,$$
  $$b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i \leq \epsilon + \xi_i^*,$$
  $$\xi_i, \xi_i^* \geq 0$$
- Let's consider the lagrange multipliers α_i, α_i^*, μ_i and μ_i^* corresponding to the above-mentioned constraints.
- The Lagrange Function is $L(\mathbf{w}, \alpha, \alpha^*, \mu, \mu^*) =$
  $$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*) + \sum_{i=1}^m \alpha_i\left(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon - \xi_i\right) +$$
  $$\sum_{i=1}^m \alpha_i^*\left(b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*\right) - \sum_{i=1}^m \mu_i\xi_i - \sum_{i=1}^m \mu_i^*\xi_i^*$$

KKT conditions for SVR

$$L(\mathbf{w}, \alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2}\|\mathbf{w}\|^2 + \cancel{C\sum_i(\xi_i + \xi_i^*)} + \sum_{i=1}^{m}\alpha_i\left(\cancel{y_i} - \mathbf{w}^\top\phi(\mathbf{x}_i) - \cancel{b - \epsilon - \xi_i}\right) +$$

$$\sum_{i=1}^{m}\alpha_i^*\left(\cancel{b +} \mathbf{w}^\top\phi(\mathbf{x}_i) - \cancel{y_i - \epsilon - \xi_i^*}\right) - \cancel{\sum_{i=1}^{m}\mu_i\xi_i - \sum_{i=1}^{m}\mu_i^*\xi_i^*}$$

- Differentiating the Lagrangian w.r.t. $\mathbf{w}$,

$$\omega + \sum_i (-\alpha_i + \alpha_i^*)\,\phi(x_i) = 0$$

$$\underbrace{\omega = \sum_i (\alpha_i - \alpha_i^*)\,\phi(x_i)}$$

Optimal ω is a linear combination of feature vectors evaluated at training data pts

KKT conditions for SVR

$$L(\mathbf{w},\alpha,\alpha^*,\mu,\mu^*) = \cancel{\frac{1}{2}\|\mathbf{w}\|^2} + C\sum_i(\xi_i + \cancel{\xi_i^*}) + \sum_{i=1}^{m}\alpha_i\left(\cancel{y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon} - \xi_i\right) + \cancel{\sum_{i=1}^{m}\alpha_i^*\left(b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*\right)} - \sum_{i=1}^{m}\mu_i\xi_i - \cancel{\sum_{i=1}^{m}\mu_i^*\xi_i^*}$$

$$\frac{\partial(x+y)}{\partial x} = 1$$

- Differentiating the Lagrangian w.r.t. $\mathbf{w}$,

$$\mathbf{w} - \alpha_i\phi(\mathbf{x}_i) + \alpha_i^*\phi(\mathbf{x}_i) = 0 \text{ i.e., } \mathbf{w} = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$$

$$\frac{\partial\left(\sum_j \xi_j\right)}{\partial \xi_i} = 1$$

- Differentiating the Lagrangian w.r.t. ξ_i, (a particular i)

$$C + (-\alpha_i) + (-\mu_i) = 0$$

i.e. $\alpha_i + \mu_i = C$

For each ξ_i one by one

KKT conditions for SVR

$$L(\mathbf{w}, \alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i (\xi_i + \xi_i^*) + \sum_{i=1}^m \alpha_i \left(y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i\right) +$$

$$\sum_{i=1}^m \alpha_i^* \left(b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*\right) - \sum_{i=1}^m \mu_i \xi_i - \sum_{i=1}^m \mu_i^* \xi_i^*$$

- Differentiating the Lagrangian w.r.t. $\mathbf{w}$,

  $$\mathbf{w} - \alpha_i \phi(\mathbf{x}_i) + \alpha_i^* \phi(\mathbf{x}_i) = 0 \text{ } i.e., \text{ } \mathbf{w} = \sum_{i=1}^m (\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$$

- Differentiating the Lagrangian w.r.t. ξ_i,
  $C - \alpha_i - \mu_i = 0$ *i.e.*, $\alpha_i + \mu_i = C$
- Differentiating the Lagrangian w.r.t ξ_i^*,
  $C - \alpha_i^* - \mu_i^* = 0$ ie $\alpha_i^* + \mu_i^* = C$

KKT conditions for SVR

$$L(\mathbf{w}, \alpha, \alpha^*, \mu, \mu^*) = \cancel{\frac{1}{2}\|\mathbf{w}\|^2} + \cancel{C\sum_i (\xi_i + \xi_i^*)} + \sum_{i=1}^m \alpha_i \left(\cancel{y_i - \mathbf{w}^\top \phi(\mathbf{x}_i)} - b \cancel{- \epsilon - \xi_i}\right) +$$
$$\sum_{i=1}^m \alpha_i^* \left(b + \cancel{\mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*}\right) - \cancel{\sum_{i=1}^m \mu_i \xi_i - \sum_{i=1}^m \mu_i^* \xi_i^*}$$

- Differentiating the Lagrangian w.r.t. $\mathbf{w}$,

  $$\mathbf{w} - \alpha_i \phi(\mathbf{x}_i) + \alpha_i^* \phi(\mathbf{x}_i) = 0 \text{ i.e., } \mathbf{w} = \sum_{i=1}^m (\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$$

- Differentiating the Lagrangian w.r.t. ξ_i,
  $C - \alpha_i - \mu_i = 0$ *i.e.*, $\alpha_i + \mu_i = C$
- Differentiating the Lagrangian w.r.t ξ_i^*,
  $\alpha_i^* + \mu_i^* = C$
- Differentiating the Lagrangian w.r.t b,

$$\frac{\partial\left(-\sum_i \alpha_i b + \sum_i \alpha_i^* b\right)}{\partial b}$$

$$-\sum_i \alpha_i + \sum_i \alpha_i^* = 0 \Rightarrow \sum_i (\alpha_i^* - \alpha_i) = 0$$

we know $\xi_i \xi_i^* = 0$ ... Do we expect $\alpha_i \alpha_i^* = 0$?

KKT conditions for SVR

$$L(\mathbf{w}, \alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i (\xi_i + \xi_i^*) + \sum_{i=1}^{m} \alpha_i \left(y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i\right) +$$
$$\sum_{i=1}^{m} \alpha_i^* \left(b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*\right) - \sum_{i=1}^{m} \mu_i \xi_i - \sum_{i=1}^{m} \mu_i^* \xi_i^*$$

- Differentiating the Lagrangian w.r.t. $\mathbf{w}$,
  $$\mathbf{w} - \alpha_i \phi(\mathbf{x}_i) + \alpha_i^* \phi(\mathbf{x}_i) = 0 \text{ i.e., } \mathbf{w} = \sum_{i=1}^{m} (\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$$
- Differentiating the Lagrangian w.r.t. ξ_i,
  $C - \alpha_i - \mu_i = 0$ *i.e.*, $\alpha_i + \mu_i = C$
- Differentiating the Lagrangian w.r.t ξ_i^*,
  $\alpha_i^* + \mu_i^* = C$
- Differentiating the Lagrangian w.r.t b,
  $\sum_i (\alpha_i^* - \alpha_i) = 0$
- Complimentary slackness: $\alpha_i (y_i - w^T \phi(x_i) - b - \epsilon - \xi_i) = 0$
  $\alpha_i^* (b + w^T \phi(x_i) - y_i - \epsilon - \xi_i^*) = 0$

At pt of optimality

KKT conditions for SVR

$$L(\mathbf{w},\alpha,\alpha^*,\mu,\mu^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i+\xi_i^*) + \sum_{i=1}^m \alpha_i\left(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon - \xi_i\right) +$$
$$\sum_{i=1}^m \alpha_i^*\left(b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*\right) - \sum_{i=1}^m \mu_i\xi_i - \sum_{i=1}^m \mu_i^*\xi_i^*$$

svm path finding algos try solving all these constraints

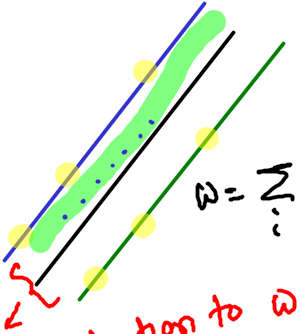
- Differentiating the Lagrangian w.r.t. $\mathbf{w}$,

$$\mathbf{w} - \alpha_i\phi(\mathbf{x}_i) + \alpha_i^*\phi(\mathbf{x}_i) = 0 \text{ i.e., } \mathbf{w} = \sum_{i=1}^m(\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$$

- Differentiating the Lagrangian w.r.t. ξ_i,
  $C - \alpha_i - \mu_i = 0$ *i.e.*, $\alpha_i + \mu_i = C$
- Differentiating the Lagrangian w.r.t ξ_i^*,
  $\alpha_i^* + \mu_i^* = C$
- Differentiating the Lagrangian w.r.t b,
  $\sum_i(\alpha_i^* - \alpha_i) = 0$
- Complimentary slackness:
  $\alpha_i(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon - \xi_i) = 0$ AND $\mu_i\xi_i = 0$ AND
  $\alpha_i^*(b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*) = 0$ AND $\mu_i^*\xi_i^* = 0$

$$\left.\begin{array}{l} y_i - w^T\phi(x_i) - b - \epsilon - \xi_i \le 0 \\ b + w^T\phi(x_i) - y_i - \epsilon - \xi_i^* \le 0 \\ \alpha_i, \alpha_i^*, \mu_i, \mu_i^* \ge 0 \\ \xi_i, \xi_i^* \ge 0 \end{array}\right\}$$

Conclusions from the KKT conditions:



No contribution to w

$$w = \sum_i (\alpha_i - \alpha_i^*)\,\phi(x_i)$$

$\mu_i \geq 0$ $\quad 0 < \alpha_i < C$ & $\alpha_i + \mu_i = C$

$\Rightarrow 0 < \mu_i < C$ & $\mu_i \xi_i = 0$

$\Rightarrow \xi_i = 0$ & $0 < \alpha_i < C$

$\alpha_i \in (0, C) \Rightarrow ?$

$\Rightarrow \xi_i = 0$ & $y_i - w^T\phi(x_i) - b - \epsilon - \cancel{\xi_i} = 0$

$\Rightarrow y_i - w^T\phi(x_i) - b = \epsilon$

$\alpha_i^* \in (0, C) \Rightarrow ?$

$w^T\phi(x_i) + b - y_i = \epsilon$

$\alpha_i = 0 \Rightarrow \mu_i = C, \xi_i = 0, y_i - w^T\phi(x_i) - b \leq \epsilon$

$y_i - w^T\phi(x_i) - b - \epsilon - \cancel{\xi_i} < 0 \Rightarrow \alpha_i = 0$ & $\mu_i = C$ & $\xi_i = 0$

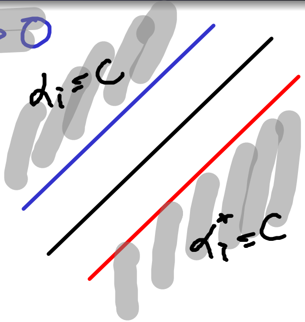
KKT conditions

- Differentiating the Lagrangian w.r.t. $\mathbf{w}$,
  $w - \alpha_i \phi(\mathbf{x}_i) + \alpha_i^* \phi(\mathbf{x}_i) = 0$
  i.e. $\mathbf{w} = \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) \phi(\mathbf{x}_i)$
- Differentiating the Lagrangian w.r.t. ξ_i,
  $C - \alpha_i - \mu_i = 0$
  i.e. $\alpha_i + \mu_i = C$
- Differentiating the Lagrangian w.r.t ξ_i^*,
  $\alpha_i^* + \mu_i^* = C$
- Differentiating the Lagrangian w.r.t b,
  $\sum_i^m (\alpha_i^* - \alpha_i) = 0$
- Complimentary slackness:
  $\alpha_i(y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i) = 0$
  $\mu_i \xi_i = 0$
  $\alpha_i^*(b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*) = 0$
  $\mu_i^* \xi_i^* = 0$

Conclusions from the KKT conditions:

If $y_i - w^\top\phi(x_i) - b - \epsilon = \xi_i > 0$

$\Rightarrow \alpha_i = C$



$$\alpha_i(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon - \xi_i) = 0$$

and

$$\alpha_i^*(b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*) = 0$$

If $b + w^\top\phi(x_i) - y_i - \epsilon = \xi_i^* > 0$

$\Rightarrow \alpha_i^* = C$

$\Rightarrow$?

Conclusions from the KKT conditions:

$$\alpha_i \in (0, C) \Rightarrow ?$$

$$(C - \alpha_i)\xi_i = 0 \Rightarrow ?$$

$$\alpha_i^* \in (0, C) \Rightarrow ?$$

$$(C - \alpha_i^*)\xi_i^* = 0 \Rightarrow ?$$

For Support Vector Regression, since the original objective and the constraints are convex, any $(\mathbf{w}, b, \alpha, \alpha^*, \mu, \mu^*, \xi, \xi^*)$ that satisfy the necessary KKT conditions gives optimality (conditions are also sufficient)

- We got some geometric intuition based on KKT

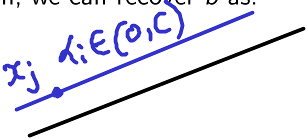
- Can we get more with some more analysis

Eg: Can we use KKT conditions to rewrite the optimization problem differently?

Some observations

- $\alpha_i, \alpha_i^* \geq 0$, $\mu_i, \mu_i^* \geq 0$, $\alpha_i + \mu_i = C$ and $\alpha_i^* + \mu_i^* = C$
  Thus, $\alpha_i, \mu_i, \alpha_i^*, \mu_i^* \in [0, C]$, $\forall i$
- If $0 < \alpha_i < C$, then $0 < \mu_i < C$
  (as $\alpha_i + \mu_i = C$)
- $\mu_i \xi_i = 0$ and $\alpha_i(y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i) = 0$ are complementary slackness conditions
  So $0 < \alpha_i < C \Rightarrow \xi_i = 0$ and $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b = \epsilon + \xi_i = \epsilon$
  - All such points lie on the boundary of the ϵ band
  - Using any point $\mathbf{x}_j$ (that is with $\alpha_j \in (0, C)$) on margin, we can recover b as:
    $b = y_j - \mathbf{w}^\top \phi(\mathbf{x}_j) - \epsilon$

$\xi_j = 0, \alpha_j > 0$



KKT Conditions, Duality, SVR Dual

KKT conditions for SVR

$$L(\mathbf{w}, \alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*) + \sum_{i=1}^m \alpha_i\left(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon - \xi_i\right) + \sum_{i=1}^m \alpha_i^*\left(b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*\right) - \sum_{i=1}^m \mu_i\xi_i - \sum_{i=1}^m \mu_i^*\xi_i^*$$

Necessary & sufficient under convexity

- Differentiating the Lagrangian w.r.t. $\mathbf{w}$,

  $$\mathbf{w} - \alpha_i\phi(\mathbf{x}_i) + \alpha_i^*\phi(\mathbf{x}_i) = 0 \text{ i.e., } \mathbf{w} = \sum_{i=1}^m(\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$$

- Differentiating the Lagrangian w.r.t. ξ_i,
  $C - \alpha_i - \mu_i = 0$ *i.e.*, $\alpha_i + \mu_i = C$
- Differentiating the Lagrangian w.r.t ξ_i^*,
  $\alpha_i^* + \mu_i^* = C$
- Differentiating the Lagrangian w.r.t b,
  $\sum_i(\alpha_i^* - \alpha_i) = 0$
- Complimentary slackness:
  $\alpha_i(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon - \xi_i) = 0$ AND $\mu_i\xi_i = 0$ AND
  $\alpha_i^*(b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*) = 0$ AND $\mu_i^*\xi_i^* = 0$

Dual problem

$\max_{\alpha_i, \alpha_i^*, \mu_i, \mu_i^* \geq 0} \; \min_{w, b, \xi_i, \xi_i^*} L(\ldots)$

Inequality becomes equality under convexity

$L^*(\alpha_i, \alpha_i^*, \mu_i, \mu_i^*)$

For Support Vector Regression, since the original objective and the constraints are convex, any $(\mathbf{w}, b, \alpha, \alpha^*, \mu, \mu^*, \xi, \xi^*)$ that satisfy the necessary KKT conditions gives optimality (conditions are also sufficient)

&

Under convexity $\max_{\alpha, \alpha^*, \mu, \mu^*, \xi, \xi^*} L^*(\alpha, \alpha^*, \mu, \mu^*, \xi, \xi^*)$

= original problem

Some observations

- $\alpha_i, \alpha_i^* \geq 0$, $\mu_i, \mu_i^* \geq 0$, $\alpha_i + \mu_i = C$ and $\alpha_i^* + \mu_i^* = C$
  Thus, $\alpha_i, \mu_i, \alpha_i^*, \mu_i^* \in [0, C]$, $\forall i$
- If $0 < \alpha_i < C$, then $0 < \mu_i < C$
  (as $\alpha_i + \mu_i = C$)
- $\mu_i \xi_i = 0$ and $\alpha_i(y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i) = 0$ are complementary slackness conditions
  So $0 < \alpha_i < C \Rightarrow \xi_i = 0$ and $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b = \epsilon + \xi_i = \epsilon$
    - All such points lie on the boundary of the ϵ band
    - Using any point $\mathbf{x}_j$ (that is with $\alpha_j \in (0, C)$) on margin, we can recover b as:
      $b = y_j - \mathbf{w}^\top \phi(\mathbf{x}_j) - \epsilon$

Support Vector Regression

Dual Objective

Weak Duality

- $L^*(\alpha, \alpha^*, \mu, \mu^*) = \min_{\mathbf{w}, b, \xi, \xi^*} L(\mathbf{w}, b, \xi, \xi^*, \alpha, \alpha^*, \mu, \mu^*)$
- By weak duality theorem, we have:
  $\min_{\mathbf{w}, b, \xi, \xi^*} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{m} (\xi_i + \xi_i^*) \geq L^*(\alpha, \alpha^*, \mu, \mu^*)$

  = under convexity

  s.t. $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b \leq \epsilon - \xi_i$, and
  $\mathbf{w}^\top \phi(\mathbf{x}_i) + b - y_i \leq \epsilon - \xi_i^*$, and
  $\xi_i, \xi^* \geq 0, \forall i = 1, \ldots, n$
- The above is true for any $\alpha_i, \alpha_i^* \geq 0$ and $\mu_i, \mu_i^* \geq 0$
- Thus,

Weak Duality

- $L^*(\alpha, \alpha^*, \mu, \mu^*) = \min_{\mathbf{w}, b, \xi, \xi^*} L(\mathbf{w}, b, \xi, \xi^*, \alpha, \alpha^*, \mu, \mu^*)$
- By weak duality theorem, we have:
  $\min_{\mathbf{w}, b, \xi, \xi^*} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m (\xi_i + \xi_i^*) \geq L^*(\alpha, \alpha^*, \mu, \mu^*)$
  s.t. $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b \leq \epsilon - \xi_i$, and
  $\mathbf{w}^\top \phi(\mathbf{x}_i) + b - y_i \leq \epsilon - \xi_i^*$, and
  $\xi_i, \xi^* \geq 0, \forall i = 1, \ldots, n$
- The above is true for any $\alpha_i, \alpha_i^* \geq 0$ and $\mu_i, \mu_i^* \geq 0$
- Thus,

$$\min_{\mathbf{w}, b, \xi, \xi^*} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m (\xi_i + \xi_i^*) \geq \max_{\alpha, \alpha^*, \mu, \mu^*} L^*(\alpha, \alpha^*, \mu, \mu^*)$$

s.t. $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b \leq \epsilon - \xi_i$, and
$\mathbf{w}^\top \phi(\mathbf{x}_i) + b - y_i \leq \epsilon - \xi_i^*$, and
$\xi_i, \xi^* \geq 0, \forall i = 1, \ldots, n$

can KKT conditions help simplify?

under convexity is equality

Dual objective

- $L^*(\alpha, \alpha^*, \mu, \mu^*) = \min_{\mathbf{w}, b, \xi, \xi^*} L(\mathbf{w}, b, \xi, \xi^*, \alpha, \alpha^*, \mu, \mu^*)$
- Assume: In case of SVR, we have a strictly convex objective and linear constraints $\Rightarrow$ KKT conditions are necessary and sufficient and strong duality holds (for $\alpha, \alpha^* \geq 0$):

$$\min_{\mathbf{w}, b, \xi, \xi^*} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{m} (\xi_i + \xi_i^*) = \max_{\alpha, \alpha^*, \mu, \mu^*} L^*(\alpha, \alpha^*, \mu, \mu^*)$$

Substitute

$\hat{\omega} = \sum_i (\alpha_i - \alpha_i^*) \phi(x_i)$

⋮

s.t. $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b \leq \epsilon - \xi_i$, and
$w^\top \phi(\mathbf{x}_i) + b - y_i \leq \epsilon - \xi_i^*$, and
$\xi_i, \xi^* \geq 0, \forall i = 1, \ldots, n$

- This value is precisely obtained at the $\left\{\hat{\mathbf{w}}, \hat{b}, \hat{\xi}, \hat{\xi}^*, \hat{\alpha}, \hat{\alpha}^*, \hat{\mu}, \hat{\mu}^*\right\}$ that satisfies the necessary (and sufficient) KKT optimality conditions [**KKT Constraint Set**]
- Given strong duality, we can equivalently solve: $\max_{\hat{\alpha}, \hat{\alpha}^*, \hat{\mu}, \hat{\mu}^*} L^*(\hat{\alpha}, \hat{\alpha}^*, \hat{\mu}, \hat{\mu}^*)$