Introduction to Machine Learning - CS725
Instructor: Prof. Ganesh Ramakrishnan
Lecture 12 - KKT Conditions, Duality, SVR Dual

KKT conditions for the Constrained (Convex) Problem

These conditions are necessary irrespective of convexity

Convexity is imp for sufficiency

- The general optimization problem we consider with (convex) inequality and (linear) equality constraints is:

$$\min_{\mathbf{w}} f(\mathbf{w})$$

$$\text{subject to } g_i(\mathbf{w}) \leq 0; 1 \leq i \leq m$$

$$h_j(\mathbf{w}) = 0; 1 \leq j \leq p$$

→ could be also represented as $h_j(\mathbf{w}) \leq 0$ & $-h_j(\mathbf{w}) \leq 0$

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

② $\nabla_w L(\hat{w}, \hat{\lambda}, \hat{\mu}) = 0$

③ $\hat{\lambda}_i g_i(\hat{w}) = 0$

s.t $\lambda_i \geq 0$

+ Other feasibility conditions $h_j(\hat{w}) = 0$ $g_i(w) \leq 0$

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:

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

Conditions:

1. $\nabla_w L(\hat{w}, \hat{\lambda}, \hat{\mu}) = 0$
2. $\hat{\lambda}_i g_i(\hat{w}) = 0$
3. $\hat{\lambda}_i \geq 0$
4. $h_j(\hat{w}) = 0$
5. $g_i(\hat{w}) \leq 0$

Necessary conditions for optimality at $\hat{w}, \hat{\lambda}, \hat{\mu}$

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:
  - $\nabla f(\hat{\mathbf{w}}) + \sum_{i=1}^{m} \hat{\lambda}_i \nabla g_i(\hat{\mathbf{w}}) + \sum_{j=1}^{p} \hat{\mu}_j \nabla h_j(\hat{\mathbf{w}}) = 0$
  - $g_i(\hat{\mathbf{w}}) \leq 0; 1 \leq i \leq m$
  - $\hat{\lambda}_i \geq 0; 1 \leq i \leq m$
  - $\hat{\lambda}_i g_i(\hat{\mathbf{w}}) = 0; 1 \leq i \leq m$
  - $h_j(\hat{\mathbf{w}}) = 0; 1 \leq j \leq p$

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:
  - $\nabla f(\hat{\mathbf{w}}) + \sum_{i=1}^{m} \hat{\lambda}_i \nabla g_i(\hat{\mathbf{w}}) + \sum_{j=1}^{p} \hat{\mu}_j \nabla h_j(\hat{\mathbf{w}}) = 0$
  - $g_i(\hat{\mathbf{w}}) \leq 0; 1 \leq i \leq m$
  - $\hat{\lambda}_i \geq 0; 1 \leq i \leq m$
  - $\hat{\lambda}_i g_i(\hat{\mathbf{w}}) = 0; 1 \leq i \leq m$
  - $h_j(\hat{\mathbf{w}}) = 0; 1 \leq j \leq p$
- 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 $(\hat{\lambda}, \hat{\mu})$

Lagrangian Duality and KKT conditions

- With $\mathbf{w} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^p$, Lagrangian is:

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

- Lagrange dual function is minimum of Lagrangian over $\mathbf{w}$.

$$L^*(\lambda, \mu) = \min_{w} L(w, \lambda, \mu)$$

Fixing values of penalty over g_i's & h_j's

Lagrangian Duality and KKT conditions

- With $\mathbf{w} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^p$, Lagrangian is:

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

- Lagrange dual function is minimum of Lagrangian over $\mathbf{w}$.

penalized lower bound

$$L^*(\lambda, \mu) = \min_{\mathbf{w}} L(\mathbf{w}, \lambda, \mu) \leq \min_{\substack{w \\ s.t\ g_i(w) \leq 0 \\ h_j(w) = 0}} f(w)$$

$\forall \lambda \geq 0$ — No λ, μ

- The Dual Optimization Problem is to maximize Lagrange dual function $L^*(\lambda, \mu)$ over (λ, μ)

Maximize the lower bound $L^*(\lambda, \mu)$ on $f(w)$

$$\max_{\substack{\lambda, \mu \\ \lambda \geq 0}} L^*(\lambda, \mu) \leq \min_{\substack{w \\ s.t\ g_i(w) \leq 0 \\ h_j(w) = 0}} f(w)$$

Lagrangian Duality and KKT conditions

- With $\mathbf{w} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^p$, Lagrangian is:

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

- Lagrange dual function is minimum of Lagrangian over $\mathbf{w}$.

$$L^*(\lambda, \mu) = \min_{\mathbf{w}} L(\mathbf{w}, \lambda, \mu)$$

- The Dual Optimization Problem is to maximize Lagrange dual function $L^*(\lambda, \mu)$ over (λ, μ)

$$\operatorname*{argmax}_{\lambda,\mu} L^*(\lambda, \mu) = \operatorname*{argmax}_{\lambda,\mu} \min_{\mathbf{w}} L(\mathbf{w}, \lambda, \mu)$$

Extra: Lagrangian Duality and KKT conditions

- The dual function yields lower bound for minimizer of the primal formulation.
- Max of dual function $L^*(\lambda, \mu)$ over (λ, μ) is also therefore a lower bound

Extra: Lagrangian Duality and KKT conditions

- The dual function yields lower bound for minimizer of the primal formulation.
- Max of dual function $L^*(\lambda, \mu)$ over (λ, μ) is also therefore a lower bound

$$\max_{\lambda,\mu} L^*(\lambda, \mu) = \max_{\lambda,\mu} \min_{\mathbf{w}} L(\mathbf{w}, \lambda, \mu) \leq L(\mathbf{w}, \lambda, \mu)$$

- **Duality Gap:** The gap between primal and dual solutions. In the KKT conditions, $\hat{\mathbf{w}}$ correspond to primal optimal and $(\hat{\lambda}, \hat{\mu})$ to dual optimal points $\Rightarrow$ Duality gap is $f(\hat{\mathbf{w}}) - L^*(\hat{\lambda}, \hat{\mu})$
- Duality gap characterizes suboptimality of the solution and can be approximated by $f(\mathbf{w}) - L^*(\lambda, \mu)$ for any feasible $\mathbf{w}$ and corresponding λ and μ

s.t $g_i(w) \leq 0$
$h_j(w) = 0$

Extra: Lagrangian Duality and KKT conditions

- The dual function yields lower bound for minimizer of the primal formulation.
- Max of dual function $L^*(\lambda, \mu)$ over (λ, μ) is also therefore a lower bound

$$\max_{\lambda,\mu} L^*(\lambda, \mu) = \max_{\lambda,\mu} \min_{\mathbf{w}} L(\mathbf{w}, \lambda, \mu) \leq L(\mathbf{w}, \lambda, \mu)$$

- **Duality Gap:** The gap between primal and dual solutions. In the KKT conditions, $\hat{\mathbf{w}}$ correspond to primal optimal and $(\hat{\lambda}, \hat{\mu})$ to dual optimal points $\Rightarrow$ Duality gap is $f(\hat{\mathbf{w}}) - L^*(\hat{\lambda}, \hat{\mu})$
- Duality gap characterizes suboptimality of the solution and can be approximated by $f(\mathbf{w}) - L^*(\lambda, \mu)$ for any feasible $\mathbf{w}$ and corresponding λ and μ
- When functions f and $g_i, \forall i \in [1, m]$ are convex and $h_j, \forall j \in [1, p]$ are affine, Karush-Kuhn-Tucker (KKT) conditions are both necessary and sufficient for points to be both primal and dual optimal with zero duality gap.

Support Vector Regression and its Dual

Instructor: Prof. Ganesh Ramakrishnan

Story: (1) Write KKT conditions

(2) Write dual optimization problem without simplifying in crudest form

(3) Simplify dual optimization problem to an equivalent formulation using KKT optimality conditions

KKT and Dual for SVR

Primal formulation: Primal vars w, ξ, ξ^*

- $\min_{\mathbf{w},b,\xi_i,\xi_i^*} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*)$
  s.t. $\forall i$,
  $y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b \leq \epsilon + \xi_i, \forall i$ (α_i)
  $b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i \leq \epsilon + \xi_i^*, \forall i$ (α_i^*)
  $\xi_i, \xi_i^* \geq 0 \ \forall i$ (μ_i, μ_i^*)

Original L_1 error SVR.

$\rightarrow -\xi_i \leq 0 \quad -\xi_i^* \leq 0$

- 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

$$L(w,\xi,\xi^*,\alpha,\alpha^*,\mu,\mu^*) = \frac{1}{2}\|w\|^2 + C\sum_i(\xi_i+\xi_i^*) + \sum_i \alpha_i(y_i - w^\top\phi(x_i) - b - \epsilon - \xi_i) + \sum_i \alpha_i^*(b + w^\top\phi(x_i) - y_i - \epsilon - \xi_i^*) - \sum_i \mu_i\xi_i - \sum_i \mu_i^*\xi_i^*$$

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

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

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

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

$$\hat{w} - \sum_i \hat{\alpha}_i \phi(x_i) + \sum_i \hat{\alpha}_i^* \phi(x_i) = 0 \quad \text{[At pt of optimality]}$$

$$\Rightarrow \hat{w} = \sum_{i=1}^m (\hat{\alpha}_i - \hat{\alpha}_i^*) \phi(x_i) \rightarrow$$ Recall similar representation for Quiz 1, prob 1

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, (for a specific "i")

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \mu_i = 0 \Rightarrow \alpha_i + \mu_i = C$$

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^*,

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

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

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. $\mathbf{w}$,
  $\mathbf{w} - \alpha_i\phi(\mathbf{x}_i) + \alpha_i^*\phi(\mathbf{x}_i) = 0$ *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:

For each i {
$\alpha_i(y_i - w^T\phi(x_i) - b - \epsilon - \xi_i) = 0$
$\alpha_i^*(b + w^T\phi(x_i) - y_i - \epsilon - \xi_i^*) = 0$
$\mu_i\xi_i = 0$ $\mu_i^*\xi_i^* = 0$

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. $\mathbf{w}$,
  $\mathbf{w} - \alpha_i\phi(\mathbf{x}_i) + \alpha_i^*\phi(\mathbf{x}_i) = 0$ *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 $\mu_i^*\xi_i^* = 0$

Conclusions from the KKT conditions:

$0 < \alpha_i < C$

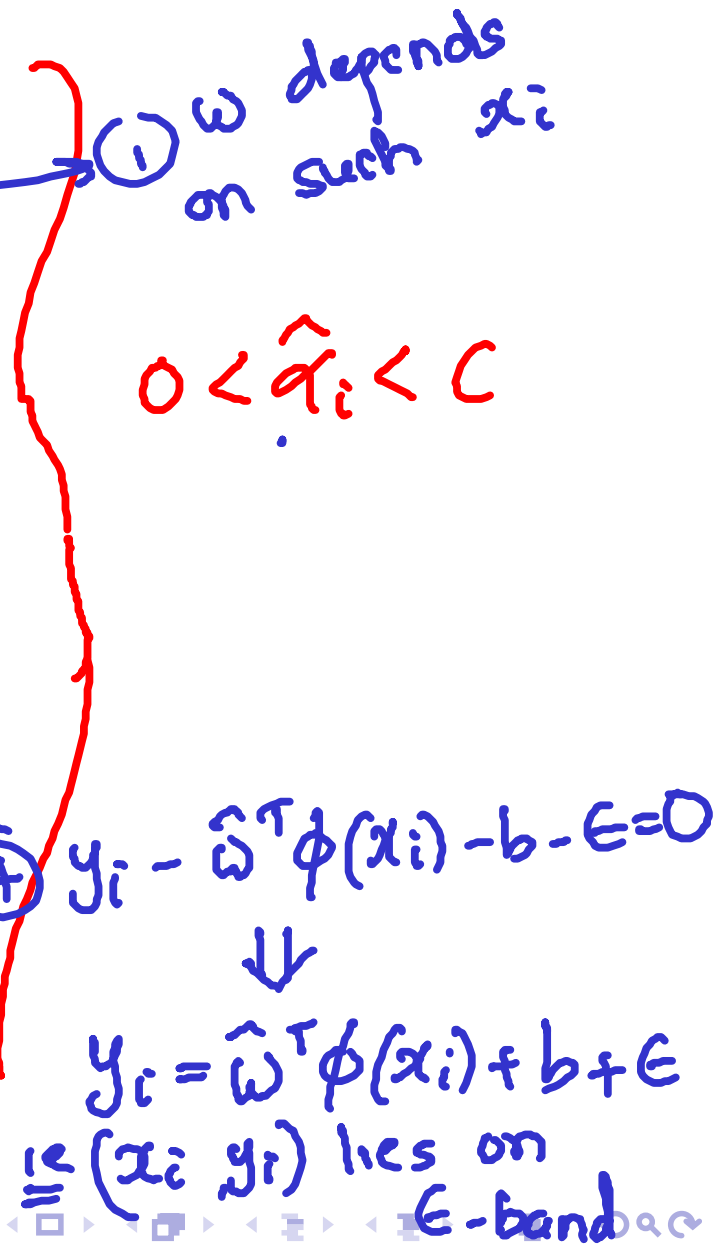
$\alpha_i \in (0, C) \Rightarrow ?$

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

$0 < \alpha_i^* < C$

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:

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

and

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

$\Rightarrow$?

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)

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:
    $b = y_j - \mathbf{w}^\top \phi(\mathbf{x}_j) - \epsilon$

In practice b is obtained as average over such x_js to cancel out numerical errors