# Introduction to Machine Learning - CS725
Instructor: Prof. Ganesh Ramakrishnan
Lecture 10 - Optimization Foundations Applied to Regression Formulations

- *Is $\nabla^2 f(\mathbf{w}^*)$ positive definite ?*
  *i.e.* $\forall \mathbf{x} \neq 0$, *is* $\mathbf{x}^T \nabla f(\mathbf{w}^*)\mathbf{x} > 0$? (A sufficient condition for local minimum)
  (Any positive definite matrix is also positive semi-definite)
  `(Cite : Section 3.12 & 3.12.1)`[1]

$$\nabla^2 f(\mathbf{w}^*) = 2\Phi^T\Phi + 2\lambda I \quad (1)$$

$$\implies \mathbf{x}^T \nabla^2 f(\mathbf{w}^*)\mathbf{x} = 2\mathbf{x}^T(\Phi^T\Phi + \lambda I)\mathbf{x} \quad (2)$$

$$= 2\left((\Phi + \sqrt{\lambda}I)\mathbf{x}\right)^T \Phi\mathbf{x} \quad (3)$$

$$= 2\left\|(\Phi + \sqrt{\lambda}I)\mathbf{x}\right\|^2 \geq 0 \quad (4)$$

- And with $\lambda = 0$, if $\Phi$ **has full column rank** ,

$$\Phi\mathbf{x} = 0 \quad \textit{iff} \quad \mathbf{x} = 0 \quad (5)$$

$\therefore$ If $\mathbf{x} \neq 0$, $\mathbf{x}^T \nabla^2 f(\mathbf{w}^*)\mathbf{x} > 0$

[1] `CS725/notes/classNotes/LinearAlgebra.pdf`

# Example of linearly correlated features

- Example where $\Phi$ doesn't have a full column rank,

$$\Phi = \begin{bmatrix} x_1 & x_1^2 & x_1^2 & x_1^3 \\ x_2 & x_2^2 & x_2^2 & x_2^3 \\ \vdots & \vdots & \vdots & \vdots \\ x_n & x_n^2 & x_n^2 & x_n^3 \end{bmatrix} \tag{6}$$

- This is the simplest form of linear correlation of features, and it is not at all desirable.
- Effect of a nonzero $\lambda$ with such $\Phi$ is that

# Example of linearly correlated features

- Example where $\Phi$ doesn't have a full column rank,

$$\Phi = \begin{bmatrix} x_1 & x_1^2 & x_1^2 & x_1^3 \\ x_2 & x_2^2 & x_2^2 & x_2^3 \\ \vdots & \vdots & \vdots & \vdots \\ x_n & x_n^2 & x_n^2 & x_n^3 \end{bmatrix} \tag{6}$$

- This is the simplest form of linear correlation of features, and it is not at all desirable.
- Effect of a nonzero $\lambda$ with such $\Phi$ is that it tends to make the Hessian more positive definite

# Do Closed-form solutions Always Exist?

- Linear regression and Ridge regression both have closed-form solutions
  - For linear regression,

$$w^* = (\Phi^\top \Phi)^{-1} \Phi^\top y$$

  - For ridge regression,

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

  (for linear regression, $\lambda = 0$)

- What about optimizing the formulations (constrained/penalized) of Lasso ($L_1$ norm)? And support-based penalty ($L_0$ norm)?: Also requires tools of Optimization/duality

# Gradient Descent Algorithm

Gradient descent is based on our previous observation that if the multivariate function $F(\mathbf{x})$ is defined and differentiable in a neighborhood of a point $\mathbf{a}$ , then $F(\mathbf{x})$ decreases fastest if one proceeds from $\mathbf{a}$ in the direction of the negative of the gradient of $F$ at $\mathbf{a}$ ,i.e. -$\nabla$ $F(\mathbf{a})$ .
Therefore,

$$\Delta \mathbf{w}^{(\mathbf{k})} = - \nabla \mathbf{E}(\mathbf{w}^{(\mathbf{k})}) \tag{7}$$

Hence,

$$\mathbf{w}^{(\mathbf{k+1})} = \mathbf{w}^{(\mathbf{k})} + \mathbf{2t}^{(\mathbf{k})}(\mathbf{\Phi}^{\mathbf{T}}\mathbf{y} - \mathbf{\Phi}^{\mathbf{T}}\mathbf{\Phi}\mathbf{w}^{(\mathbf{k})} - \lambda \mathbf{w}^{(\mathbf{k})}) \tag{8}$$

# Gradient Descent Algorithm

**Find** starting point $\mathbf{w^{(0)}} \epsilon \mathcal{D}$

- $\Delta \mathbf{w^k} = -\nabla \varepsilon(\mathbf{w^{(k)}})$
- Choose a step size $t^{(k)} > 0$ using exact or backtracking ray search.
- Obtain $\mathbf{w^{(k+1)}} = \mathbf{w^{(k)}} + \mathbf{t^{(k)}} \boldsymbol{\Delta} \mathbf{w^{(k)}}$.
- Set $k = k + 1$. **until** stopping criterion (such as $\|\nabla \varepsilon(\mathbf{w}^{(k+1)}) \| \leq \epsilon$) is satisfied

**Exact line search algorithm to find $t^{(k)}$**

- The line search approach first finds a descent direction along which the objective function f will be reduced and then computes a step size that determines how far **x** should move along that direction.
- In general,

$$t^{(k)} = \arg\min_t f\left(\mathbf{w}^{(\mathbf{k+1})}\right) \tag{9}$$

- Thus,

# Gradient Descent Algorithm

**Exact line search algorithm to find $t^{(k)}$**

- The line search approach first finds a descent direction along which the objective function f will be reduced and then computes a step size that determines how far **x** should move along that direction.
- In general,

$$t^{(k)} = \arg\min_{t} f\left(\mathbf{w^{(k+1)}}\right) \tag{9}$$

- Thus,

$$t^{(k)} = \arg\min_{t} \left(\mathbf{w^{(k)}} + \mathbf{2t}\left(\mathbf{\Phi^{T}y} - \mathbf{\Phi^{T}}\phi\mathbf{w^{(k)}} - \lambda\mathbf{w^{(k)}}\right)\right) \tag{10}$$

# Example of Gradient Descent Algorithm

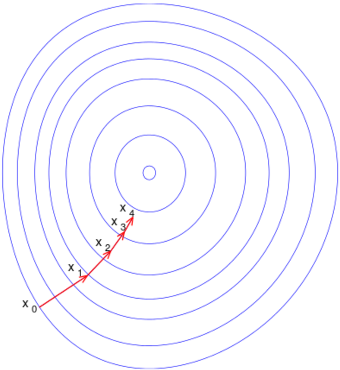



Figure 1: A red arrow originating at a point shows the direction of the negative gradient at that point. Note that the (negative) gradient at a point is orthogonal to the level curve going through that point. We see that gradient descent leads us to the bottom of the bowl, that is, to the point where the value of the function F is minimal. Source: Wikipidea

# Constrained Least Squares Linear Regression

Find

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \|\phi\mathbf{w} - \mathbf{y}\|^2 \;\; s.t. \;\; \|\mathbf{w}\|_p \leq \zeta, \tag{11}$$

where

$$\|\mathbf{w}\|_p = \left(\sum_{i=1}^{n} |w_i|^p\right)^{\frac{1}{p}} \tag{12}$$

**Claim: This is an equivalent reformulation of the penalized least squares. Why?**

# p-Norm level curves

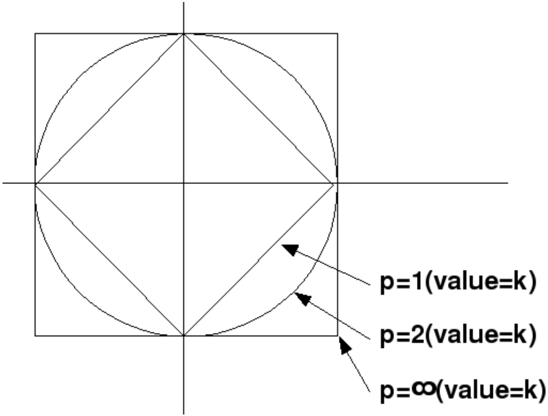



Figure 2: p-Norm curves for constant norm value and different p

# Convex Optimization Problem

- Formally, a convex optimization problem is an optimization problem of the form

$$\textit{minimize } f(\mathbf{x}) \tag{13}$$
$$\textit{subject to } c \in C \tag{14}$$

  where f is a convex function, C is a convex set, and **x** is the optimization variable.

- An improved form of the above would be

$$\textit{minimize } f(\mathbf{x}) \tag{15}$$
$$\textit{subject to } g_i(\mathbf{x}) \leq 0, \; i = 1, ..., m \tag{16}$$
$$h_i(\mathbf{x}) = 0, i = 1, ..., p \tag{17}$$

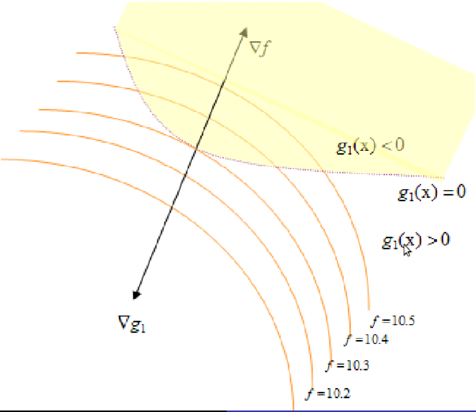
  where f is a convex function, $g_i$ are convex functions, and $h_i$ are affine functions, and **x** is the vector of optimization variables.

# Constrained convex problems

**Q.** *How to solve constrained problems of the above-mentioned type?*

**A.** General problem format :

$$\textit{Minimize } f(\mathbf{w}) \; s.t. \; g(\mathbf{w}) \leq 0 \tag{18}$$



# Constrained Convex Problems

- At the point of optimality,

$$\textit{Either } g(\mathbf{w}^*) < 0 \quad \& \quad \nabla f(\mathbf{w}^*) = 0 \tag{19}$$

$$\textit{Or } g(\mathbf{w}^*) = 0 \quad \& \quad \nabla f(\mathbf{w}^*) = \alpha \nabla g(\mathbf{w}^*) \tag{20}$$

- If $\mathbf{w}^*$ is on the boundary of $g$, *i.e.*, $g(\mathbf{w}^*) = 0$,

$$\nabla f(\mathbf{w}^*) = \alpha \nabla g(\mathbf{w}^*) \tag{21}$$

*(Duality Theory)* `(Cite : Section 4.4, pg-72)`[2]

[2] `cs709/notes/BasicsOfConvexOptimization.pdf`

# Constrained Convex Problems

- At the point of optimality,

$$\textit{Either } g(\mathbf{w}^*) < 0 \quad \& \quad \nabla f(\mathbf{w}^*) = 0 \tag{19}$$

$$\textit{Or } g(\mathbf{w}^*) = 0 \quad \& \quad \nabla f(\mathbf{w}^*) = \alpha \nabla g(\mathbf{w}^*) \tag{20}$$

- If $\mathbf{w}^*$ is on the boundary of $g$, *i.e.*, $g(\mathbf{w}^*) = 0$,

$$\nabla f(\mathbf{w}^*) = \alpha \nabla g(\mathbf{w}^*) \tag{21}$$

*(Duality Theory)* `(Cite : Section 4.4, pg-72)`[2]

- **Intuition:** If the above didn't hold, then we would have $\nabla f(\mathbf{w}^*) = \alpha_1 \nabla g(\mathbf{w}^*) + \alpha_2 \nabla_\perp g(\mathbf{w}^*)$, where by moving in direction $\pm \nabla_\perp g(\mathbf{w}^*)$, we remain on boundary $g(\mathbf{w}^*) = 0$, while decreasing/increasing value of f, which is not possible at the point of optimality.

[2] `cs709/notes/BasicsOfConvexOptimization.pdf`

# "Regularized" Linear Regression

- We limit the weights of the coefficients by putting a constraint on size of the L2 norm of the weight vector

$$\arg\min_{\mathbf{w}} (\mathbf{\Phi w} - \mathbf{Y})^T (\mathbf{\Phi w} - \mathbf{Y})$$

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

- 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?)

# Duality and KKT conditions

For a convex objective and constraint function, the minima, $\mathbf{w}^*$, can satisfy one of the following two conditions:

1. $g(\mathbf{w}^*) = \mathbf{0}$ and $\nabla f(\mathbf{w}^*) = \alpha \nabla \mathbf{g}(\mathbf{w}^*)$
2. $g(\mathbf{w}^*) < \mathbf{0}$ and $\nabla f(\mathbf{w}^*) = \mathbf{0}$

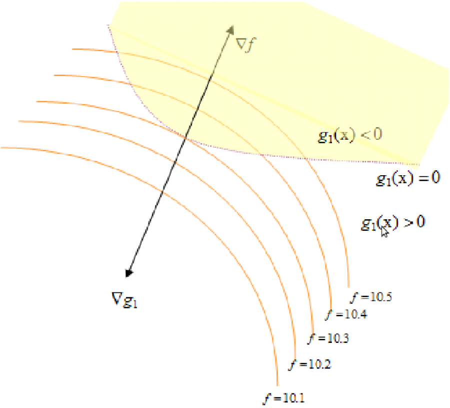



Figure 4: Two conditions when a minima can occur: a) When the minima is on the constraint function boundary, in which case the gradients are along the same direction ;b) When minima is inside the constraint space (shown in yellow shade), in which case $\nabla f(\mathbf{w}^*) = \mathbf{0}$.

# Duality and KKT conditions

- This fact can be easily visualized from the previous figure. As we can see, the first condition occurs when minima lies on the boundary of function $g$. In this case, gradient vectors corresponding to the function $f$ and the function $g$, at $\mathbf{w}^*$, point in the same direction barring multiplication by a real constant.
- Second condition depicts the case when minima lies inside the constraint space. This space is shown shaded in Figure 1. Clearly, for this case $\nabla f(\mathbf{w}) = \mathbf{0}$ for minima to occur. This primal problem can be converted to dual using the lagrange multiplier. According to which, we can convert this problem to the objective function augmented by weighted sum of constraint functions in order to get the corresponding lagrangian.

$$L(\mathbf{w}, \lambda) = \mathbf{f}(\mathbf{w}) + \lambda \mathbf{g}(\mathbf{w}); \lambda \in \mathbb{R}$$

# Duality and KKT conditions

- Here, we wish to penalize higher magnitude coefficients, hence, we wish $g(\mathbf{w})$ to be negative while minimizing the lagrangian. In order to maintain such direction, we must have $\lambda \geq 0$. Also, for solution $\mathbf{w}$ to be feasible, $\nabla g(\mathbf{w}) \leq \mathbf{0}$.
- Due to complementary slackness condition, we further have $\lambda g(\mathbf{w}) = \mathbf{0}$, which roughly suggests that the lagrange multiplier is zero unless constraint is active at the minimum point. As $\mathbf{w}$ minimizes the lagrangian $L(\mathbf{w}, \lambda)$, gradient must vanish at this point and hence we have $f(\mathbf{w}) + \lambda \nabla \mathbf{g}(\mathbf{w}) = \mathbf{0}$

# Duality and KKT conditions

- In general, optimization problem with inequality and equality constraints might be depicted in the following manner:

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

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

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

# Duality and KKT conditions

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

$$L(\mathbf{w}, \lambda, \mu) = \mathbf{f}(\mathbf{w}) + \sum_{\mathbf{i}=1}^{\mathbf{m}} \lambda_{\mathbf{i}} \mathbf{g}_{\mathbf{i}}(\mathbf{w}) + \sum_{\mathbf{j}=1}^{\mathbf{p}} \mu_{\mathbf{j}} \mathbf{h}_{\mathbf{j}}(\mathbf{w})$$

- Lagrange dual function is the minimum value of the lagrangian over $\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^p$.

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

# Duality and KKT conditions

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

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

- The gap between primal and dual solutions is the duality gap,
- Duality gap characterizes suboptimality of the solution.

$$f(\mathbf{w}) - \mathbf{L}^*(\lambda, \mu)$$

- 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.

# Duality and KKT conditions

For above mentioned formulation of the problem, KKT conditions for all differentiable functions (i.e. $f, g_i, h_j$) with $\hat{\mathbf{w}}$ primal optimal and $(\hat{\lambda}, \hat{\mu})$ dual optimal point are:

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

# Bound on $\lambda$ in the regularized least square solution

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

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

(the first KKT condition). Here, $f(\mathbf{w}) = (\phi\mathbf{w} - \mathbf{Y})^{\mathbf{T}}(\phi\mathbf{w} - \mathbf{Y})$ and, $g(\mathbf{w}) = \|\mathbf{w}\|^{\mathbf{2}} - \xi$.
Solving we get,

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

From the second KKT condition we get,

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

From the third KKT condition,

$$\lambda \geq 0$$

From the fourth condition

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

# Bound on $\lambda$ in the regularized least square solution

Values of **w** and $\lambda$ that satisfy all these equations would yield an optimal solution. 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,

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

Using the triangle inequality we obtain,

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

# Bound on $\lambda$ in the regularized least square solution

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

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

i.e.

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

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

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

This is not the exact solution of $\lambda$ but the bound proves the existence of $\lambda$ for some $\xi$ and $\phi$.

# 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 $\lambda$. This form of **regularized** regression is often referred to as **Ridge regression**.

# Support Vector Regression and its Dual

Instructor: Prof. Ganesh Ramakrishnan

# KKT and Dual for SVR

- $\min_{w,b,\xi_i,\xi_i^*} \frac{1}{2}\|w\|^2 + C\sum_i(\xi_i + \xi_i^*)$
  s.t. $\forall i$,
  $y_i - w^\top \phi(x_i) - b \leq \epsilon + \xi_i$,
  $b + w^\top \phi(x_i) - y_i \leq \epsilon + \xi_i^*$,
  $\xi_i, \xi_i^* \geq 0$
- Let's consider the lagrange multipliers $\alpha_i$, $\alpha_i^*$, $\mu_i$ and $\mu_i^*$ corresponding to the above-mentioned constraints respectively.

# KKT conditions

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

# Conclusions from the KKT conditions:

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

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

# KKT conditions

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

# Conclusions from the KKT conditions:

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

and

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

$\Rightarrow$ ?

# 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 $(\mathbf{w}, \mathbf{b}, \alpha, \alpha^*, \mu, \mu^*, \xi, \xi^*)$ 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 $\alpha_i(y_i - -w^\top \phi(x_i) - b - \epsilon - \xi_i) = 0$ are complementary slackness conditions
  So $0 < \alpha_i < C \Rightarrow \xi_i = 0$ and $y_i - w^\top \phi(x_i) - b = \epsilon + \xi_i = \epsilon$
  - All such points lie on the boundary of the $\epsilon$ band
  - Using any point $x_j$ (that is with $\alpha_j \in (0, C)$) on margin, we can recover $b$ as:
    $b = y_j - w^\top \phi(x_j) - \epsilon$

# Support Vector Regression

## Dual Objective

# Dual function

- Let $L^*(\alpha, \alpha^*, \mu, \mu^*) = \min_{w,b,\xi,\xi^*} L(w, b, \xi, \xi^*, \alpha, \alpha^*, \mu, \mu^*)$
- By weak duality theorem, we have:
  $\min_{w,b,\xi,\xi^*} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n}(\xi_i + \xi_i^*) \geq L^*(\alpha, \alpha^*, \mu, \mu^*)$
  s.t. $y_i - w^\top \phi(x_i) - b \leq \epsilon - \xi_i$, and
  $w^\top \phi(x_i) + b - y_i \leq \epsilon - \xi_i^*$, 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,

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

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

# Dual objective

- In case of Support Vector Regression, we have a strictly convex objective and linear constraints $\Rightarrow$ KKT conditions are necessary and sufficient and strong duality holds:

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

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

- This value is precisely obtained at the $(\mathbf{w}, \mathbf{b}, \xi, \xi^*, \alpha, \alpha^*, \mu, \mu^*)$ that satisfies the necessary (and sufficient) optimality conditions
- Given strong duality, we can equivalently solve

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

- $L(\alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) + \sum_{i=1}^{n}\left(\alpha_i(y_i - w^\top\phi(x_i) - b - \epsilon - \xi_i) + \alpha_i^*(w^\top\phi(x_i) + b - y_i - \epsilon - \xi_i^*)\right) \sum_{i=1}^{n}(\mu_i\xi_i + \mu_i^*\xi_i^*)$
- We obtain $w$, $b$, $\xi_i$, $\xi_i^*$ in terms of $\alpha$, $\alpha^*$, $\mu$ and $\mu^*$ by using the KKT conditions derived earlier as $w = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\phi(x_i)$ and $\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0$ and $\alpha_i + \mu_i = C$ and $\alpha_i^* + \mu_i^* = C$
- Thus, we get:
$$L(w, b, \xi, \xi^*, \alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2}\sum_i\sum_j(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\phi^\top(x_i)\phi(x_j) + \sum_i\left(\xi_i(C - \alpha_i - \mu_i) + \xi_i^*(C - \alpha_i^* - \mu_i^*)\right) - b\sum_i(\alpha_i - \alpha_i^*) - \epsilon\sum_i(\alpha_i + \alpha_i^*) + \sum_i y_i(\alpha_i - \alpha_i^*) - \sum_i\sum_j(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\phi^\top(x_i)\phi(x_j)$$
$$= -\frac{1}{2}\sum_i\sum_j(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\phi^\top(x_i)\phi(x_j) - \epsilon\sum_i(\alpha_i + \alpha_i^*) + \sum_i y_i(\alpha_i - \alpha_i^*)$$

# Kernel function: $K(x_i, x_j) = \phi^T(x_i)\phi(x_j)$

- $w = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\phi(x_i) \Rightarrow$ the final decision function $f(x) = w^T\phi(x) + b = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\phi^T(x_i)\phi(x) + y_j - \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\phi^T(x_i)\phi(x_j) - \epsilon$
  $x_j$ is any point with $\alpha_j \in (0, C)$
- The dual optimization problem to compute the $\alpha$'s for SVR is:

$$max_{\alpha_i, \alpha_i^*} - \frac{1}{2}\sum_i \sum_j (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\phi^\top(x_i)\phi(x_j)$$

$$-\epsilon \sum_i (\alpha_i + \alpha_i^*) + \sum_i y_i(\alpha_i - \alpha_i^*)$$

  s.t.
  - $\sum_i (\alpha_i - \alpha_i^*) = 0$
  - $\alpha_i, \alpha_i^* \in [0, C]$
- **We notice that the only way these three expressions involve $\phi$ is through $\phi^\top(x_i)\phi(x_j) = K(x_i, x_j)$, for some $i, j$**