# Introduction to Machine Learning - CS725

Instructor: Prof. Ganesh Ramakrishnan

## Lecture 10 - Optimization Foundations Applied to Regression Formulations

# Building on questions on Least Squares Linear Regression

1. Is there a probabilistic interpretation?
   - Gaussian Error, Maximum Likelihood Estimate
2. Addressing overfitting
   - Bayesian and Maximum Aposteriori Estimates, Regularization, Support Vector Regression
3. How to minimize the resultant and more complex error functions?
   - Level Curves and Surfaces, Gradient Vector, Directional Derivative, Gradient Descent Algorithm, Convexity, Necessary and Sufficient Conditions for Optimality

# Constrained convex problems

**Q.** *How to solve such constrained problems?*

**A.** Canonical example:

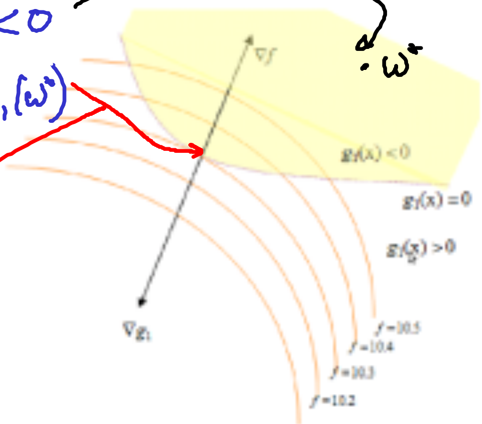
$$\textit{Minimize } f(\mathbf{w}) \ s.t. \ g_1(\mathbf{w}) \leq 0 \tag{1}$$

At $w^*$ (optimal pt)

$\nabla f(w^*)$ & $g_1(w^*) < 0$

<u>OR</u> $\nabla f(w^*) = -\lambda \nabla g_1(w^*)$

$g_1(w^*) = 0$



# Constrained Convex Problems

- If $\mathbf{w}^*$ is on the boundary of $g_1$, *i.e.*, if $g_1(\mathbf{w}^*) = 0$,

$$\nabla f(\mathbf{w}^*) = -\lambda \nabla g_1(\mathbf{w}^*) \text{ for some } \lambda \geq 0$$

- **Intuition:** If the above didn't hold, then we would have $\nabla f(\mathbf{w}^*) = \lambda_1 \nabla g_1(\mathbf{w}^*) + \lambda_2 \nabla_\perp g_1(\mathbf{w}^*)$, where, by moving in direction[1] $\pm\nabla_\perp g_1(\mathbf{w}^*)$ ( or $-\nabla g_1(\mathbf{w}^*)$), we remain on boundary $g_1(\mathbf{w}^*) = 0$, ( or within $g_1(\mathbf{w}^*) \leq 0$) while decreasing the value of $f$, which is not possible at the point of optimality.
- Thus, at the point of optimality[2], for some $\lambda \geq 0$,

$$\text{Either } g_1(\mathbf{w}^*) < 0 \quad \& \quad \nabla f(\mathbf{w}^*) = 0 \;\; (\lambda = 0) \tag{2}$$

$$\text{Or } g_1(\mathbf{w}^*) = 0 \quad \& \quad \nabla f(\mathbf{w}^*) = -\lambda \nabla g_1(\mathbf{w}^*) \tag{3}$$

Summary:

$L(\omega) = f(\omega) + \lambda g_1(\omega)$

$\nabla L(\omega^*) = 0$

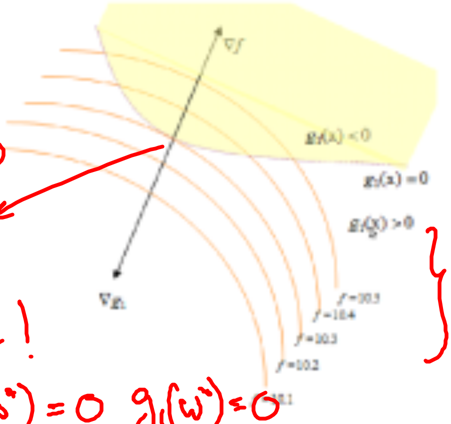
Comp. Slackness $\leftarrow$ $\lambda g_1(\omega^*) = 0$

$g_1(\omega^*) \leq 0$

[1] $\nabla_\perp g_1(\mathbf{w}^*)$ is the direction orthogonal to $\nabla g_1(\mathbf{w}^*)$

[2] Section 4.4, pg-72: cs725/notes/BasicsOfConvexOptimization.pdf

# Explaining the Figure



If $\lambda = 0$ & $g_1(w^*) = 0$

$\Rightarrow$ Constraint $g_1(w^*) \leq 0$ is not playing any role!

$\nabla f(w^*) = 0 \quad g_1(w^*) = 0$

$\underset{}{\operatorname{argmax}}\ f(w)$ s.t. $g_1(w) \leq 0$

$||$

$\operatorname{argmax}\ f(w)$

Figure 2: Two conditions under which a minimum can occur: a) When the minimum is on the constraint function boundary, in which case the gradients are in opposite directions; b) When point of minimum is inside the constraint space (shown in yellow shade), in which case $\nabla f(\mathbf{w}^*) = \mathbf{0}$.

# More Explanation and Lagrange Function

- The first condition occurs when minima lies on the boundary of function $g$. In this case, gradient vectors corresponding to the functions $f$ and $g$, at $\mathbf{w}^*$, point in opposite directions barring multiplication by a real constant.
- Second condition represents the case that point of minimum lies inside the constraint space. This space is shown shaded in Figure 1. Clearly, for this case, $\nabla f(\mathbf{w}) = \mathbf{0}$.
- An Alternative Representation: $\nabla L(\mathbf{w}, \lambda) = 0$ for some $\lambda \geq 0$ where

original objective

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

Penalized constraint

is called the lagrange function which has objective function augmented by weighted sum of constraint functions

# 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 $\nabla f(\mathbf{w}^*) = -\lambda \nabla \mathbf{g}(\mathbf{w}^*)$
2. $g(\mathbf{w}^*) < \mathbf{0}$ and $\nabla f(\mathbf{w}^*) = \mathbf{0}$

# 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 $\nabla f(\mathbf{w}) + \lambda \nabla \mathbf{g}(\mathbf{w}) = \mathbf{0}$

# KKT Conditions, Duality, SVR Dual

# KKT conditions for the Constrained (Convex) Problem

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

Illustrate through ← Quadratic fn (SVR)

$$\|Aw - b\|_2^2$$

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

# KKT conditions for the Constrained (Convex) Problem

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

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

Set of inequalities
Eg: linear $Bw \geq c$

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

Set of equalities
Eg: linear $Dw = f$

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

# 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:

Goal: Minimize $L$

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

Penalty associated with each inequality $g_i(w) \le 0$

Diff penalty with each equality

$\Downarrow$

Intuition (Necessary at opt)

① Since we need $g_i(w) \le 0$, $\lambda_i \ge 0$

$$\nabla f(w^*) = -\underbrace{\sum_i \lambda_i \nabla g_i(w^*)} - \sum_j \mu_j \nabla h_j(w^*)$$

② Since we need $h_j(w) = 0$, sign of $\mu_j$ does not matter

$h_j(w^*) = 0$

$g_i(w^*) \le 0$

$\lambda_i g_i(w^*) = 0$

# KKT conditions for the Constrained (Convex) Problem

Karush Kuhn Tucker

- 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:
  1. $\nabla f(\hat{\mathbf{w}}) + \sum_{i=1}^{m} \hat{\lambda}_i \nabla g_i(\hat{\mathbf{w}}) + \sum_{j=1}^{p} \hat{\mu}_j \nabla h_j(\hat{\mathbf{w}}) = 0$ Gradient equality
  2. $g_i(\hat{\mathbf{w}}) \leq 0; 1 \leq i \leq m$ $\rightarrow$ ineq (original)
  3. $\hat{\lambda}_i \geq 0; 1 \leq i \leq m$ $\rightarrow$ positivity
  4. $\hat{\lambda}_i g_i(\hat{\mathbf{w}}) = 0; 1 \leq i \leq m$ $\rightarrow$ comp slackness
  5. $h_j(\hat{\mathbf{w}}) = 0; 1 \leq j \leq p$ $\rightarrow$ eq (original)

Let us apply to constrained ridge regression

See pages 23 to 29 for elaboration

$$\min_{w} \|\phi w - y\|_2^2$$

$$\text{s.t } \|w\|_2^2 \le \theta \rightarrow \lambda$$

Looks like soln to penalized ridge regression

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

$$L(\omega, \lambda) = \|\phi\omega - y\|_2^2 + \lambda\left(\|\omega\|_2^2 - \theta\right)$$

① $\nabla L(\hat{\omega}, \hat{\lambda}) = 0 \Rightarrow 2\Phi^T\Phi\hat{\omega} - 2\Phi^T y + 2\lambda\hat{\omega} = 0$

② $\|\hat{\omega}\|_2^2 \le \theta$ ③ $\hat{\lambda} \ge 0$ ④ $\hat{\lambda}\left(\|\hat{\omega}\|_2^2 - \theta\right) = 0$ $\Big\}$ $\hat{\lambda} = f(\theta)$

$\hat{\lambda}$ s.t.

$\hat{\lambda}$ st if $\|(\phi^T\phi)^{-1}\Phi^T y\|_2^2 \le \Theta$ then $\hat{\lambda} = 0$ else smallest $\hat{\lambda} > 0$

s.t $\|(\phi^T\phi + \hat{\lambda} I)^{-1}\phi^T y\|_2^2 = \Theta$

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

$$\left.\begin{array}{l} h_j \leq 0 \\ -h_j \leq 0 \end{array}\right\} \text{both convex}$$

ie 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 $(\hat{\lambda}, \hat{\mu})$

# 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 $\mathbf{w}$.

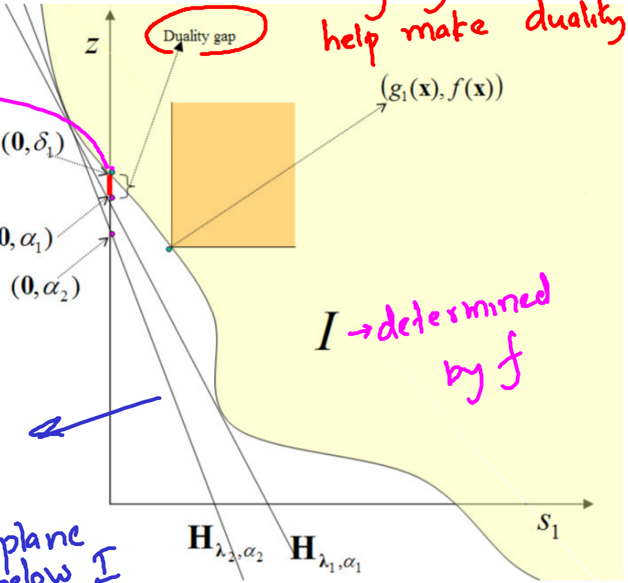
$$L^*(\lambda, \mu) = \min_{w} L(w, \lambda, \mu) \leq f(w) \quad \text{s.t} \quad g_i(w) \leq 0, \; h_j(w) = 0$$

$$\underbrace{\max_{\lambda, \mu} L^*(\lambda, \mu) \qquad\qquad}_{} \leq f(w) \quad \text{s.t} \quad g_i(w) \leq 0, \; h_j(w) = 0$$

Red gap in image will always lie above Hyperplane

Convexity of I would help make duality gap → 0

y intercept of I ≡ optimal value of f subject to constraints




I → determined by f

$L^*(\lambda, \mu)$ corresponds to a Hyperplane below I

# 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 $(\lambda, \mu)$

$\hat{\lambda}, \hat{\mu} = \arg\max_{\lambda, \mu} L^*(\lambda, \mu) \equiv$ Push the Hyperplane as upward as possible

- 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 $(\lambda, \mu)$

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

tends to = for convex prob.

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

$g_i(\omega) \leq 0$

$h_j(\omega) = 0$

# 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 $(\lambda, \mu)$ is also therefore a lower bound

$$\underbrace{L^*(\lambda,\mu)}_{\text{max over } \lambda,\mu} \le \underbrace{f(w)}_{\text{min over } w} \quad \forall\, w, \lambda, \mu$$

$\lambda \geq 0$     s.t. $g_i(w) \le 0$, $h_j(w) = 0$

# 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 $(\lambda, \mu)$ 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 $\lambda$ and $\mu$

If $\hat{\lambda}, \hat{\mu}, \hat{w}$ is soln to KKT condition then $f(\hat{w}) - L^*(\hat{\lambda}, \hat{\mu})$ is gap.

$= 0$ under convexity

Most imp

# 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 $(\lambda, \mu)$ 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 $\lambda$ and $\mu$
- 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.

Elaboration on equivalence of penalized & constrained forms of ridge regression (continued from page 14)

# 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 $\mathbf{w}^*$

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

(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

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

# Equivalent Forms of Ridge Regression

- Values of $\mathbf{w}$ and $\lambda$ that satisfy all these equations would yield an optimal solution. That is, if
$$\|\mathbf{w}^*\| = \|(\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}\| \leq \xi$$
then $\lambda = 0$ is the solution. Else, for some sufficiently large value, $\lambda$ will be the solution to
$$\|\mathbf{w}^*\| = \|(\Phi^T\Phi + \lambda I)^{-1}\Phi^T\mathbf{y}\| = \xi$$

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

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

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 $\alpha = \|(\Phi^T\Phi)\|$. Substituting in the previous equation,

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

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

# 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,

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

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 $\lambda$ but the bound proves the existence of $\lambda$ for some $\xi$ and $\Phi$.

# 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 $\lambda$. This form of **regularized** ridge regression is the **penalized ridge regression**.