Introduction to Machine Learning - CS725
Instructor: Prof. Ganesh Ramakrishnan
Lecture 10 - Optimization for Regression and Machine Learning Concluded

Recap: Iterative Soft Thresholding Algorithm for Solving Lasso

Recap: Proximal Subgradient Descent for Lasso

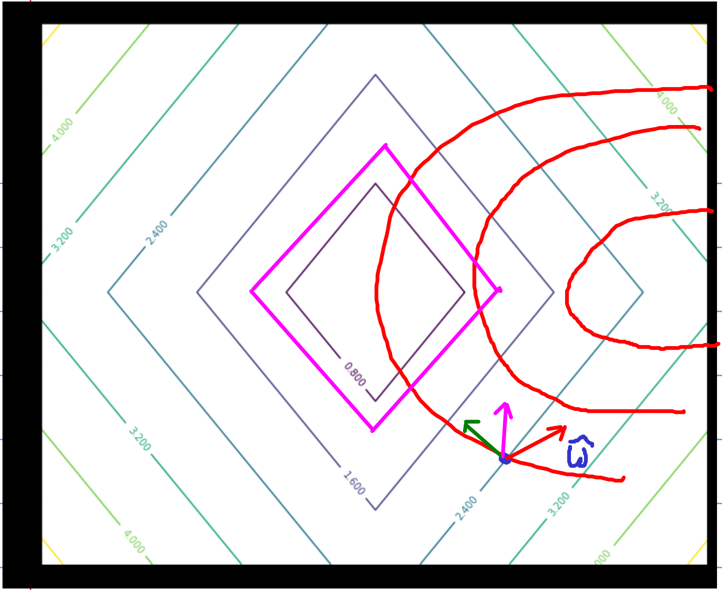
- Let $\varepsilon(\mathbf{w}) = \|\phi\mathbf{w} - \mathbf{y}\|_2^2$
- **Proximal Subgradient Descent Algorithm:**
  Initialization: Find starting point $\mathbf{w}^{(0)}$
  - Let $\widehat{\mathbf{w}}^{(\mathbf{k+1})}$ be a next gradient descent iterate for $\varepsilon(\mathbf{w}^k)$
  - Compute $\mathbf{w}^{(k+1)} = \underset{\mathbf{w}}{\mathbf{argmin}}||\mathbf{w} - \widehat{\mathbf{w}}^{(\mathbf{k+1})}||_{\mathbf{2}}^{\mathbf{2}} + \lambda\mathbf{t}||\mathbf{w}||_{\mathbf{1}}$ by setting subgradient of this objective to $\mathbf{0}$. This results in (see https://www.cse.iitb.ac.in/~cs725/notes/classNotes/lassoElaboration.pdf)
    1. ...
    2. ...
    3. ...
  - Set $k = k+1$, **until** stopping criterion is satisfied (such as no significant changes in $\mathbf{w}^k$ w.r.t $\mathbf{w}^{(k-1)}$)

Recap: Iterative Soft Thresholding Algorithm (Proximal Subgradient Descent) for Lasso

- Let $\varepsilon(\mathbf{w}) = \|\phi\mathbf{w} - \mathbf{y}\|_2^2$
- **Iterative Soft Thresholding Algorithm:**
  Initialization: Find starting point $\mathbf{w}^{(0)}$
  - Let $\widehat{\mathbf{w}}^{(\mathbf{k+1})}$ be a next iterate for $\varepsilon(\mathbf{w}^k)$ computed using using any (gradient) descent algorithm
  - Compute $\mathbf{w}^{(k+1)} = \underset{\mathbf{w}}{argmin}||\mathbf{w} - \widehat{\mathbf{w}}^{(\mathbf{k+1})}||_2^2 + \lambda\mathbf{t}||\mathbf{w}||_1$ by:
    1. If $\widehat{w}_i^{(k+1)} > \lambda t/2$, then $w_i^{(k+1)} = -\lambda t/2 + \widehat{w}_i^{(k+1)}$
    2. If $\widehat{w}_i^{(k+1)} < -\lambda t/2$, then $w_i^{(k+1)} = \lambda t/2 + \widehat{w}_i^{(k+1)}$
    3. 0 otherwise. → $\hat{w}_i^{(k+1)} \in \left[-\frac{\lambda t}{2}, \frac{\lambda t}{2}\right]$ are set to 0
  - Set $k = k + 1$, **until** stopping criterion is satisfied (such as no significant changes in $\mathbf{w}^k$ w.r.t $\mathbf{w}^{(k-1)}$)

Cause of sparsity

larger the val of λ, more the # of zero w_i's


0.800
1.600
2.400
3.200
4.000
$\hat{\omega}$

Dealing with Constraints in Optimization

Recap: Constrained Least Squares Linear Regression

Find

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \|\phi\mathbf{w} - \mathbf{y}\|^2 \;\; s.t. \;\; \|\mathbf{w}\|_p \le \zeta, \tag{1}$$

where

$$\|\mathbf{w}\|_p = \left(\sum_{i=1}^{n} |w_i|^p\right)^{\frac{1}{p}} \tag{2}$$

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

p-Norm level curves

Expect the green vector (linear combination of ↘ ↙ to characterize solutions to the constrained problem

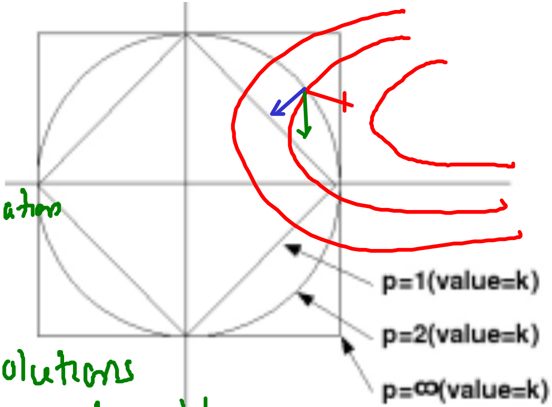


Figure 1: p-Norm curves for constant norm value and different p

Recap: SVR objectives (Rewriting each inequality as ≤ 0)

- 1-norm Error, and L_2 regularized:
  - $\min_{\mathbf{w},b,\xi_i,\xi_i^*} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*)$ $\quad f(w, \xi_i, \xi_i^*)$
    s.t. $\forall i$,
    $y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b \leq \epsilon + \xi_i$, $\rightarrow g_i(w, \xi_i, \xi_i^*) = y_i - w^T\phi(x_i) - b - \epsilon - \xi_i \leq 0$
    $b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i \leq \epsilon + \xi_i^*$, $\rightarrow g_i^*(w, \xi_i, \xi_i^*) = b + w^T\phi(x_i) - y_i - \epsilon - \xi_i^* \leq 0$
    $\xi_i, \xi_i^* \geq 0$ $\quad -\xi_i, -\xi_i^* \leq 0$
- 2-norm Error, and L_2 regularized:
  - $\min_{\mathbf{w},b,\xi_i,\xi_i^*} \frac{1}{2}\|w\|^2 + C\sum_i(\xi_i^2 + \xi_i^{*2})$
    s.t. $\forall i$,
    $y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b \leq \epsilon + \xi_i$,
    $b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i \leq \epsilon + \xi_i^*$
  - Here, the constraints $\xi_i, \xi_i^* \geq 0$ are not necessary

Convex Optimization Problem

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

$$
\begin{aligned}
minimize\ f(\mathbf{w}) \qquad (3)\\
subject\ to\ c \in C \qquad (4)
\end{aligned}
$$

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

- A specific form of the above would be

$$
\begin{aligned}
&minimize\ f(\mathbf{w}) && && (5)\\
&subject\ to\ g_i(\mathbf{w}) &\leq\ & 0,\ i = 1, \ldots, m && (6)\\
&h_i(\mathbf{w}) &=\ & 0, i = 1, \ldots, p && (7)
\end{aligned}
$$

Constrained convex problems

Q. *How to solve such constrained problems?*

A. Canonical example:

$$\textit{Minimize } f(\mathbf{w}) \text{ s.t. } g_1(\mathbf{w}) \leq 0 \tag{8}$$

Case 2: Constraint holds naturally without need for imposing

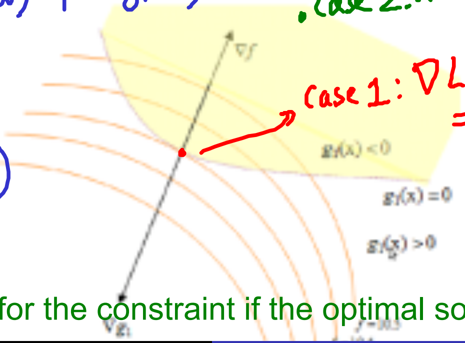
$L(\omega, \lambda) = f(\omega) + \lambda g_1(\omega)$

$\nabla L(\omega, \lambda)$

$= \nabla f(\omega) + \lambda \nabla g(\omega)$

case 2: $\lambda = 0$, $\nabla L = 0$ $\Rightarrow \nabla f = 0$

case 1: $\nabla L = 0$ $\Rightarrow \nabla f = -\lambda \nabla g$



Case 2:I do not care for the constraint if the optimal solution already satisifies it

Constrained Convex Problems

(case 1)

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

$\nabla L(w^*)=0 \Rightarrow$
$$\nabla f(\mathbf{w}^*) = -\lambda \nabla g_1(\mathbf{w}^*) \text{ for some } \lambda \geq 0,$$
else Case 2

- **Intuition:** See `https://lvnext.lokavidya.com/courses/1/modules/items/162`
- At the point of optimality[1], for some $\lambda \geq 0$,

How do we avoid "either"

case 1

$$\boxed{\begin{array}{ll} \textit{Either } g_1(\mathbf{w}^*) < 0 \;\&\; \nabla f(\mathbf{w}^*) = 0 \quad (\lambda = 0) & (9) \\ \textit{Or } g_1(\mathbf{w}^*) = 0 \;\&\; \nabla f(\mathbf{w}^*) = -\lambda \nabla g_1(\mathbf{w}^*) & (10) \end{array}}$$

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

Case 2

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

The Lagrange Function

- At the point of optimality, for some $\lambda \geq 0$,

$$\begin{aligned} \textit{Either } g_1(\mathbf{w}^*) < 0 \;\; \& \;\; \nabla f(\mathbf{w}^*) = 0 \qquad (11) \\ \textit{Or } g_1(\mathbf{w}^*) = 0 \;\; \& \;\; \nabla f(\mathbf{w}^*) = -\lambda \nabla g_1(\mathbf{w}^*) \qquad (12) \end{aligned}$$

- An Alternative Representation: $\underline{\nabla L(\mathbf{w}, \lambda) =}\ 0$ for some $\lambda \geq 0$ where

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

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 $\nabla f(\mathbf{w}^*) = \mathbf{0}$

Combine using

$x=0$ or $y=0$

iff

$x.y=0$

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:

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

KKT conditions for the Constrained (Convex) Problem

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

-h & h are convex when h is linear

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

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

$h_j(w) \leq 0$

$-h_j(w) \leq 0$

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

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

$$\nabla L(w^*, \lambda, \mu) = 0$$

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:
  $L(\mathbf{w}, \lambda, \mu) = f(\mathbf{w}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{w}) + \sum_{j=1}^{p} \mu_j h_j(\mathbf{w})$
- 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 $\nabla f(\hat{\mathbf{w}}) + \sum_{i=1}^{m} \hat{\lambda}_i \nabla g_i(\hat{\mathbf{w}}) + \sum_{j=1}^{p} \hat{\mu}_j \nabla h_j(\hat{\mathbf{w}}) = 0$
  - 4 $g_i(\hat{\mathbf{w}}) \leq 0; 1 \leq i \leq m$ and $h_j(\hat{\mathbf{w}}) = 0; 1 \leq j \leq p$ 5
  - $\underbrace{\hat{\lambda}_i \geq}_{3} 0; 1 \leq i \leq m$ and $\underbrace{\hat{\lambda}_i g_i}_{2}(\hat{\mathbf{w}}) = 0; 1 \leq i \leq m$

Numbering as per what was written on the black board in class

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

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

Interested in behaviour of L in (λ, μ) space after getting rid of ω.

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 **w**.

maximize

$$L^*(\lambda, \mu) = \min_{\mathbf{w}} L(\mathbf{w}, \lambda, \mu) \leq \text{soln to Original prob}$$

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

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)

becomes "=" under convexity

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

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

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

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 $\mathbf{w}^*$

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

(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,

Equivalent Forms of Ridge Regression

- Values of $\mathbf{w}$ and λ that satisfy all these equations would yield an optimal solution. That is, if

Case 2 (Interior)

$$\left\{ \|\mathbf{w}^*\|^2 = \|(\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}\|^2 \leq \xi \right. \to \text{with } \lambda = 0$$

(soln in interior of constraint)

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

Case 1 (boundary)

$$\left\{ \begin{array}{l} \|\mathbf{w}^*\|^2 = \|(\Phi^T\Phi + \lambda I)^{-1}\Phi^T\mathbf{y}\|^2 = \xi \\ \|w^*\|^2 = \xi \end{array} \right.$$

[As $\lambda \uparrow$ $\|w^*\|^2 \uparrow$]

Bound on λ 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,

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

Using the triangle inequality we obtain, $\|a\| + \|b\| \geq \|a+b\|$

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

- 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,

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

known known

i.e.

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.

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

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

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

RHS is all in terms of knowns

This is not the exact solution of λ but the bound proves the

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

No fun without "Duality"