Lecture 07- Convex and Constrained Optimization and Regression

Instructor: Prof. Ganesh Ramakrishnan

Agenda

- Overfitting
- Regularized Regression and Constrained Convex Optimization
- Support Vector Regression

Redundant ϕ and Overfitting

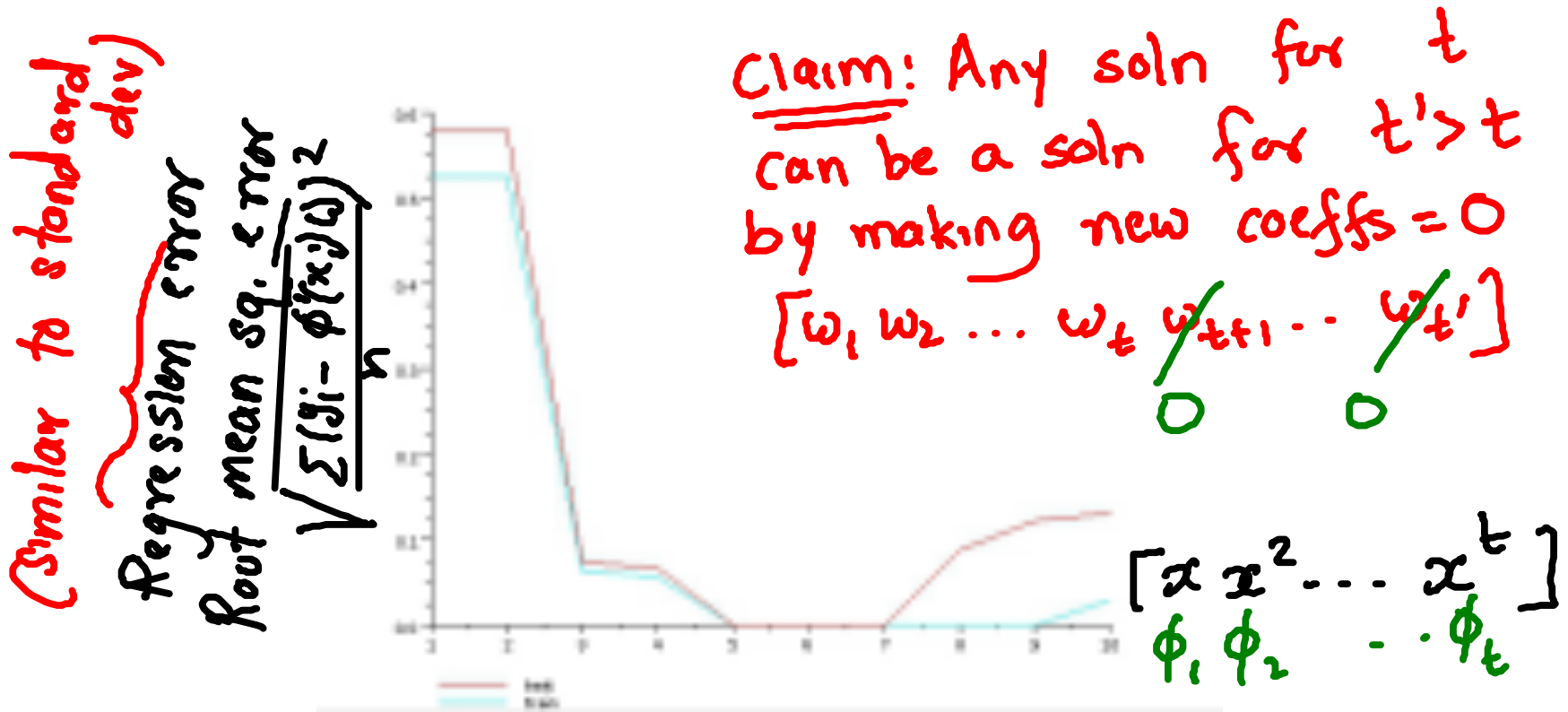


Figure: train-RMS and test-RMS values vs t(degree of polynomial) graph

- Too many bends (t=9 onwards) in curve $\equiv$ high values of some $w_i s$
- Train and test errors differ significantly

Constrained Least Squares Linear Regression

Find

$$\mathbf{w}^* = \text{argmin}_{\mathbf{w}} \|\phi\mathbf{w} - \mathbf{y}\|^2 \;\; s.t. \;\; \|\mathbf{w}\|_p^p \leq \zeta, \tag{1}$$

$\Omega(w) \leq \zeta$

where

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

Why? Suppose $p \to 0$, Note: $x^0 = 1$ except if $x = 0$

$0^p = 0$

$\lim_{p \to 0} \|w\|_p$ = # of non-zero w_i's

Also called support/cardinality of w

eg of feature selection

For other p's let us look at level curves of $\|w\|_p$

p-Norm level curves

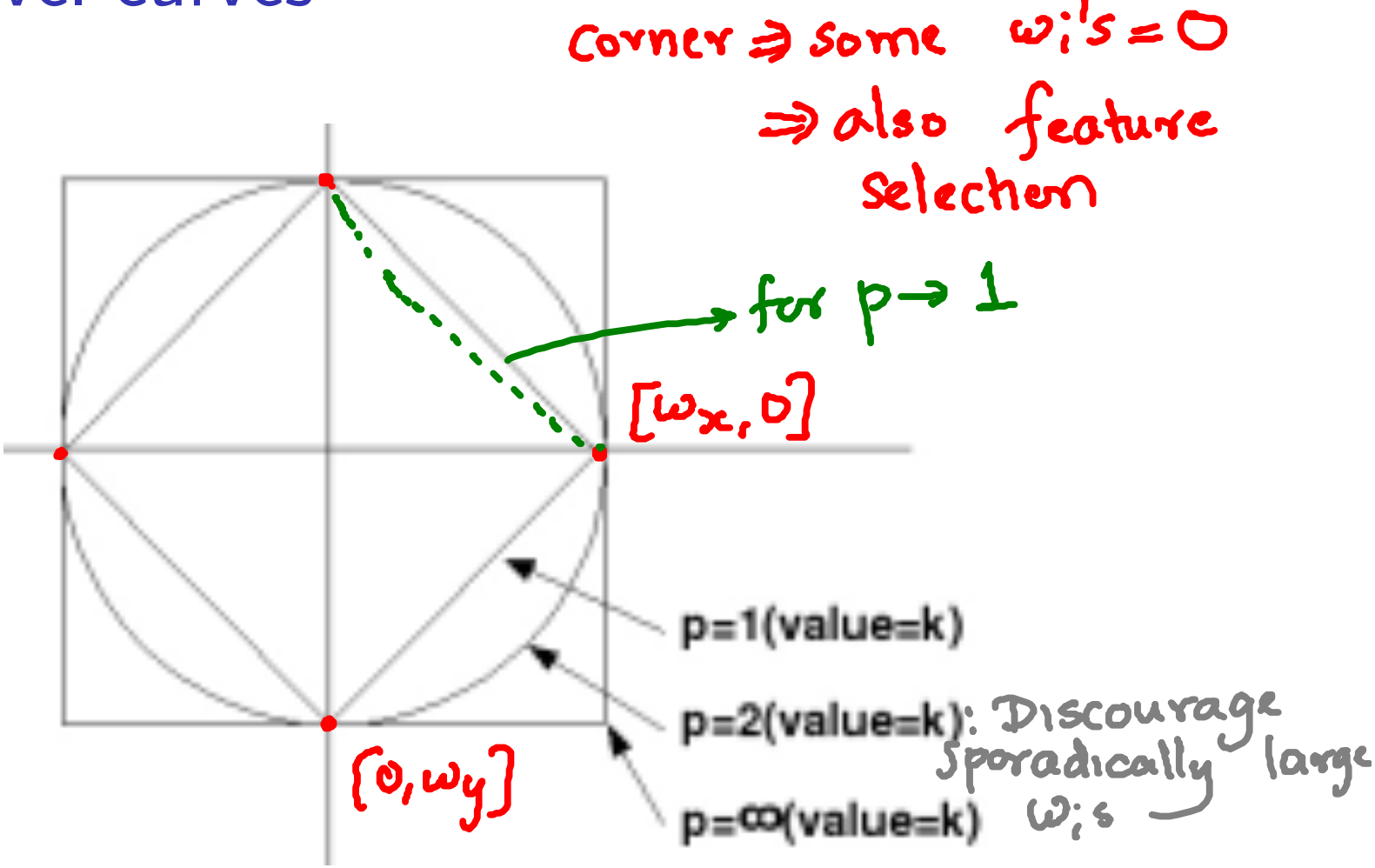


Figure: p-Norm curves for constant norm value and different p

Convex Optimization Problem

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

$$\textit{minimize } f(\mathbf{x}) \tag{3}$$
$$\textit{subject to } c \in C \tag{4}$$

where f is a convex function, C is a convex set, and $\mathbf{x}$ is the optimization variable.

- An improved form of the above would be

$$\textit{minimize } f(\mathbf{x}) \tag{5}$$
$$\textit{subject to } g_i(\mathbf{x}) \leq 0,\ i = 1, ..., m \tag{6}$$
$$h_i(\mathbf{x}) = 0, i = 1, ..., p \tag{7}$$

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

Constrained convex problems

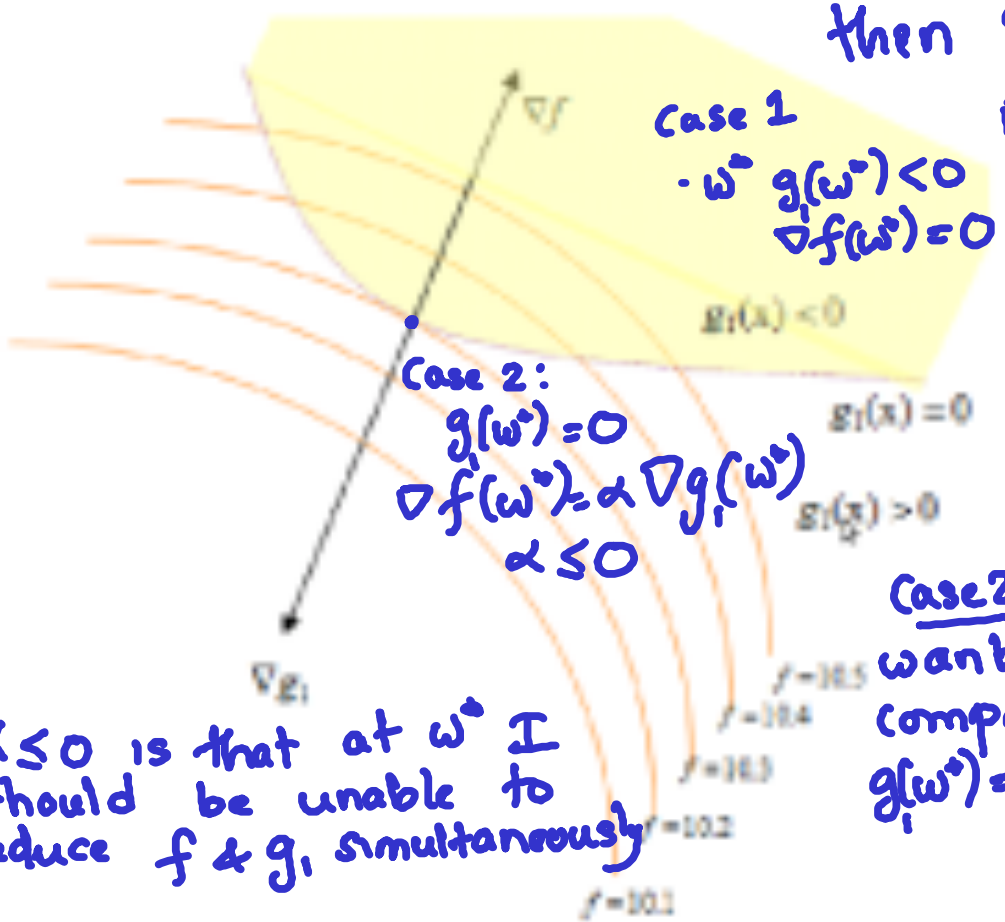
Q. *How to solve constrained problems of the above-mentioned type?*

A. General problem format :

$$\textit{Minimize } f(\mathbf{w}) \textit{ s.t. } g(\mathbf{w}) \leq 0 \quad (8)$$

$= f(w^*)$

If $g(w)$ is a convex fn then $\{w | g(w) \leq 0\}$ is a convex set



Case 1
- w^* $g(w^*) < 0$
  $\nabla f(w^*) = 0$

Case 2:
$g(w^*) = 0$
$\nabla f(w^*) = \alpha \nabla g_1(w^*)$
$\alpha \leq 0$

Case2: We do not want ∇f to have component along $g(w^*) = 0$

$\alpha \leq 0$ is that at w^* I should be unable to reduce f & g_1 simultaneously

Constrained Convex Problems

- At the point of optimality,

$$\textit{Either } g(\mathbf{w}^*) < 0 \quad \& \quad \nabla f(\mathbf{w}^*) = 0 \tag{9}$$

$$\textit{Or } g(\mathbf{w}^*) = 0 \quad \& \quad \nabla f(\mathbf{w}^*) = \alpha \nabla g(\mathbf{w}^*) \tag{10}$$

($\alpha \leq 0$)

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

α is a Lagrange multiplier

$$\nabla f(\mathbf{w}^*) = \alpha \nabla g(\mathbf{w}^*) \tag{11}$$

(Duality Theory) (Cite : Section 4.4, pg-72)[1]

- **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.

[1] 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

$$\text{argmin}_{\boldsymbol{w}}(\boldsymbol{\Phi w} - \boldsymbol{Y})^T(\boldsymbol{\Phi w} - \boldsymbol{Y})$$

$g_1(\mathbf{w}) = \|\mathbf{w}\|_2^2 - \xi \leq 0 \Leftrightarrow \|\mathbf{w}\|_2^2 \leq \xi$

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

Require α s.t $\|(\Phi^T\Phi - 2\alpha I)^{-1}\Phi^T y\|_2^2 = \xi$

i.e $\|w^*\|_2^2 = \xi$ & $(\Phi^T\Phi)w^* - \Phi^T y = 2\alpha w^* \Rightarrow w^* = (\Phi^T\Phi - 2\alpha I)^{-1}\Phi^T y$

OR $\|w^*\|_2^2 < \xi$ & $(\Phi^T\Phi)w^* - \Phi^T y = 0 \Rightarrow w^* = (\Phi^T\Phi)^T\Phi^T y$

Requires: $\|(\Phi^T\Phi)^{-1}\Phi^T y\|_2^2 < \xi$

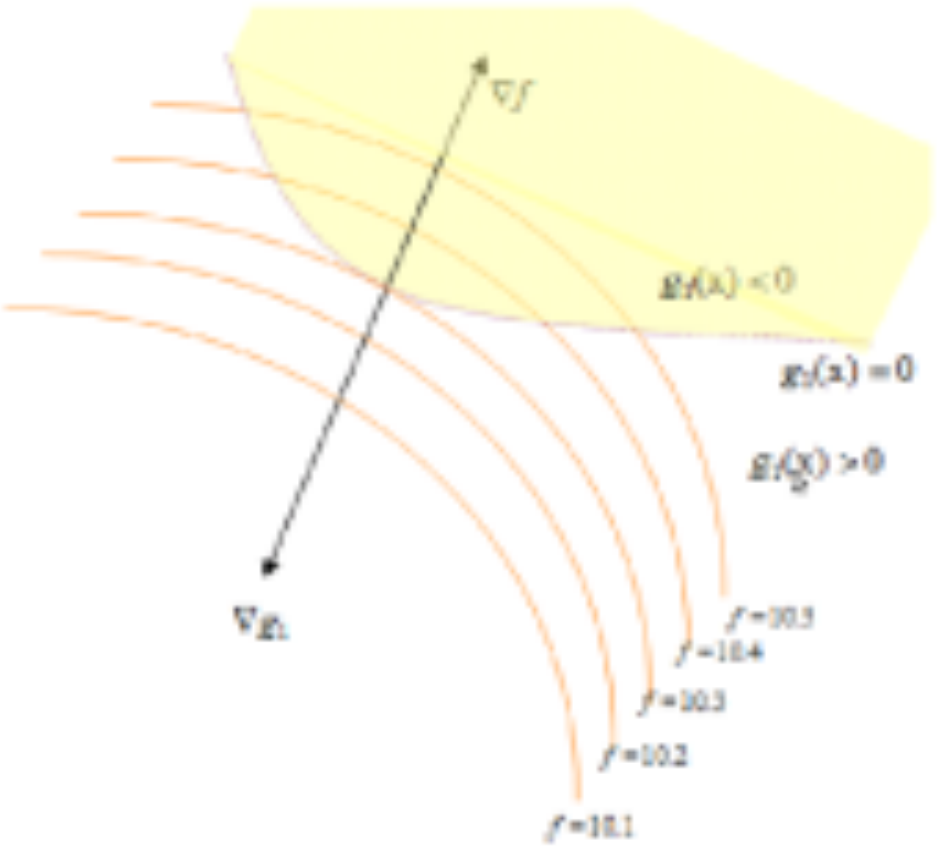


Figure: Two conditions when a minima can occur: a) When the minima is on the constraint function boundary, in which case the gradients are along the same direction ;b) When minima is inside the constraint space (shown in yellow shade), in which case $\nabla f(\mathbf{w}^*) = \mathbf{0}$.

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:

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

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

inequality

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

equality

Duality and KKT conditions

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

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

$\lambda_i \geq 0$

Multipliers

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

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

Necessary condition for $\min_w L(w,\lambda,\mu)$ at $\hat{w}$: $\nabla L(\hat{w}) = 0 \Rightarrow \nabla f(\hat{w}) + \sum \lambda_i \nabla g_i(\hat{w}) + \sum \mu_j \nabla h_j(\hat{w}) = 0$

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

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

- 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 may be given in the following manner:

$$\nabla_{\mathbf{w}}\left(f(\hat{\mathbf{w}}) + \sum_{i=1}^{m} \hat{\lambda}_i \nabla g_i(\hat{\mathbf{w}}) + \sum_{j=1}^{p} \hat{\mu}_j h_j(\hat{\mathbf{w}})\right) = 0$$

- $g_i(\hat{\mathbf{w}}) \leq 0; 1 \leq i \leq m$
- $\hat{\lambda}_i \geq 0; 1 \leq i \leq m$
- $\hat{\lambda}_i g_i(\hat{\mathbf{w}}) = 0; 1 \leq i \leq m$
- $h_j(\hat{\mathbf{w}}) = 0; 1 \leq j \leq p$

If $\hat{w}$ on bndry $\Rightarrow g_i(\hat{w}) = 0$

If $\hat{w}$ NOT on bndry $\Rightarrow \nabla f(\hat{w}) = 0$

$\Leftarrow \nabla f(\hat{w}) = \underbrace{0}_{\lambda_i = 0} \cdot \nabla g_i(\hat{w})$

Complementary slackness condition