LECTURE - 19 : 13th Feb.

When H is symmetric, $\text{Null}(H) = \text{Null}(H^T)$

Hence $\text{Null}(H) \perp \text{Range}(H)$.

Now let us examine

$$\underline{\min_{x \in \mathbb{R}^n} \underline{\tfrac{1}{2} x^T H x + c^T x + d}}, \quad H \text{ is symmetric.}$$

Case 1 : $\underline{H \text{ is positive definite}}$ → unique sol^n

Case 2 : H is positive semidefinite with one of its eigenvalues $= 0$ → unbounded $\Leftrightarrow c \notin R(H)$.

Case 3 : H has a negative eigenvalue. → unbounded

→ let λ be an eigenvalue that is $-ve$, and u be the corresponding eigenvector.

$$\bar{x} = \alpha u, \quad \alpha \in \mathbb{R}$$

the function value
$$\frac{1}{2} \alpha^2 u^T H u + c^T \alpha u + d$$
$$= \frac{1}{2} \alpha^2 u^T (\lambda u) + c^T \alpha u + d$$
$$= \frac{\lambda}{2} \alpha^2 \underbrace{u^T u}_{+ve} + \underline{c^T \alpha u} + d$$

$-ve$.

as $\alpha \to \infty$, the first term will go to $-\infty$

$\Rightarrow f(\alpha u) \to -\infty$ as $\alpha \to \infty$.

$\Rightarrow$ the problem is unbounded and does not have a finite optimal value.

Case 1 : $\nabla f(x) = Hx + c$, the optimal solution x^* satisfies

$\nabla^2 f(x) = H \succ 0$

$$Hx^* + c = 0$$

$$\Rightarrow \underline{x^* = -H^{-1} c} \; : \text{unique optimal sol}^n.$$

Case 2: $H \succeq 0$ with at least one eigenvalue $\lambda = 0$.

$\nabla^2 f(x) = H \succeq 0$.

necessary condition for optimality is $Hx^* + c = 0$

$\Rightarrow Hx^* = -c$

Case 2a: $\boxed{c \notin \text{Range}(H)} \Rightarrow \nexists\ \bar{x}$ satisfying $H\bar{x} = -c$.

Since $\lambda = 0$ is an eigenvalue,

let $x = \alpha u$ where u is the eigenvector of $\lambda = 0$.

$$f(\alpha u) = \frac{1}{2}\alpha^2 \underbrace{u^T H u}_{=0} + \frac{1}{2}\alpha c^T u + d = \frac{1}{2}\alpha c^T u + d.$$

HW · show that $c^T u \neq 0$.

Since $c^T u \neq 0$, we can choose $\alpha \to \infty$ or $\alpha \to -\infty$ s.t.

$f(\alpha u) \to -\infty$. Optimal value is not finite & the problem is unbounded.

Case 2b: $c \in \text{Range}(H) \Rightarrow \exists\ \bar{x}$ satisfying $H\bar{x} = -c$

then any $\bar{x} + \alpha u$ where $u \in \text{Null}(H)$

also satisfies $H(\bar{x} + \alpha u) = H\bar{x} + \alpha Hu = -c$

Let us compute

$$f(\bar{x} + \alpha u) = \frac{1}{2}(\bar{x} + \alpha u)^T H (\bar{x} + \alpha u) + c^T(\bar{x} + \alpha u) + d$$

$$= \frac{1}{2}\bar{x}^T H \bar{x} + c^T \bar{x} + \alpha c^T u + d$$

$[c^T u = (-H\bar{x})^T u = (\bar{x})^T H u = 0]$.

$= \frac{1}{2}\bar{x}^T H \bar{x} + c^T \bar{x} + d$ irrespective of choice of α and u.

$v^T H^T H v = \|Hv\|_2^2 \geq 0$

Now recall that the linear regression problem $\underset{w \in \mathbb{R}^K}{\min} \| \phi(\hat{x}) w - \hat{y} \|_2^2 =: f(\omega)$

$$f(\omega) = (\phi(\hat{x})\omega - \hat{y})^T(\phi(\hat{x})\omega - \hat{y}) = \omega^T \phi(\hat{x})^T \phi(\hat{x})\omega - 2\omega^T \phi(\hat{x})^T \hat{y} + (\hat{y})^T \hat{y}$$

$\phi : \mathbb{R}^n \to \mathbb{R}^K$

$\phi(\hat{x})^T \phi(\hat{x}) \in \mathbb{R}^{K \times K}$.

$$= \frac{1}{2} x^T H x + c^T x + d,$$

where $x = \omega$, $H = 2\phi(\hat{x})^T \phi(\hat{x})$

$d = (\hat{y})^T \hat{y}$

$c = -2\phi(\hat{x})^T \hat{y}$

If $H \succ 0$, we have a unique optimal solution $H\omega^* = -c$

$$\Rightarrow \phi(\hat{x})^T \phi(\hat{x}) \omega^* = \phi(\hat{x})^T \hat{y}$$

$$\Rightarrow y^* = [\phi(\hat{x})^T \phi(\hat{x})]^{-1} \phi(\hat{x})^T \hat{y}.$$

If H has an eigenvalue $= 0$, we can show that

$\phi(\hat{x})^T \hat{y} \in \text{Range}(H)$.

and we have infinitely many solutions, with the same finite optimal value.

When the number of features $K >$ number of data points,

$\phi(\hat{x}) \in \mathbb{R}^{N \times K}$ has fewer rows that columns.

$\text{rank}(\phi(\hat{x})) < K \Rightarrow$ there are infinitely many solutions,

$\bar{\omega} + \eta v$, $v \in N(\phi)$.

When a new data point $\tilde{x}$ is given, then

$$\phi(\tilde{x})(\bar{\omega}) \neq \phi(\tilde{x})(\bar{\omega} + \eta v)$$

In fact, the entries of $\bar{\omega}$ and $\bar{\omega} + \eta v$ are going to be very different from each other, and it is difficult to predict y on new data points.

In order to alleviate the above issue, the regression problem is modified to include regularization terms, either in the cost function or as constraints.

Ex:

$$\min_{w} \|\phi(\hat{x})w - \hat{y}\|_2^2 + \lambda \|w\|_2^2 \quad , \quad \lambda : \text{hyperparameter}$$

→ cost function is now strongly convex & we have unique optimal solution.

Sparsity promoting solutions:

card(w) = number of non-zero entries of the vector

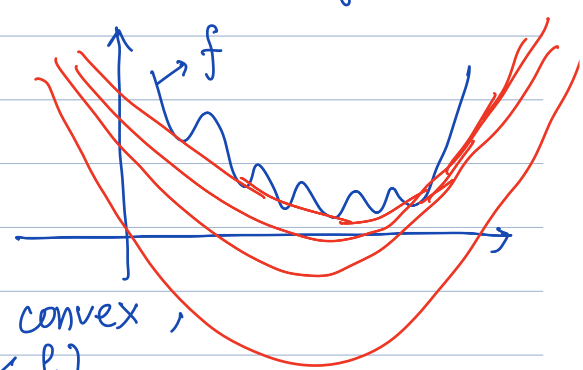
$$= \left|\{ i \mid w_i \neq 0 \}\right|$$

$$\left[\begin{array}{l} \min_{w} \|\phi(\hat{x})w - \hat{y}\|_2^2 \\ \text{s.t.} \quad \text{card}(w) \leq m \end{array}\right]$$

→ NP-Hard problem as card(w) is not a convex function.

we can approximate the above by a convex optimization problem.

For a function f, its (convex) envelope



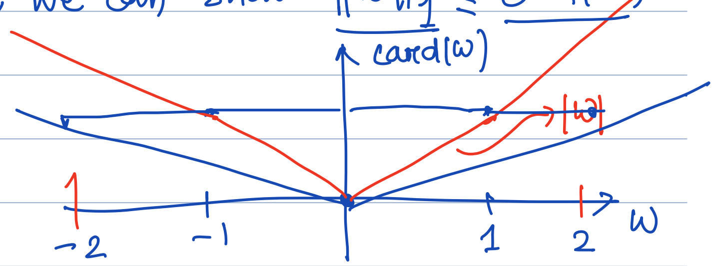
$$\text{env} f = \sup\{\phi : \phi \text{ is convex}, \ \phi \leq f\}$$

Under certain assumptions, we can show $\|w\|_1 \leq \text{card}(w)$.

Ex: let $w \in \mathbb{R}$,

$w \in [-1, 1]$

$|w| \leq \text{card}(w)$



$\{w \mid \|w\|_\infty \leq R\}$ card(w) $\geq \frac{1}{R}\|w\|_1$

Lecture 20: 14th Feb.

Motivated by the above observation, we define ℓ_1-regularized regression problem as

$$\min_{w \in \mathbb{R}^K} \|\phi(\hat{x})w - \hat{y}\|_2^2 + \lambda \|w\|_1 .$$

an alternative formulation

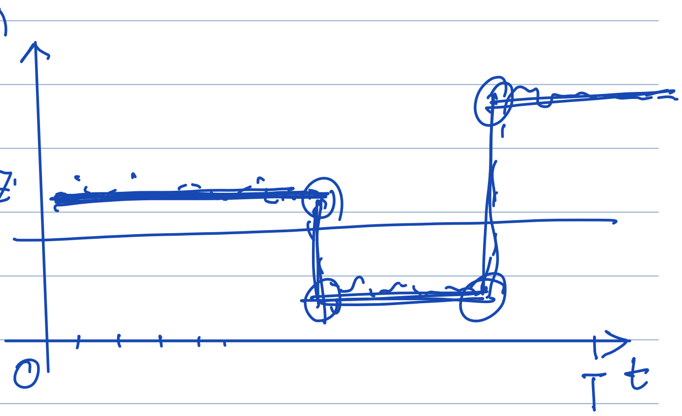
$$\min_{w \in \mathbb{R}^K} \|\phi(\hat{x})w - \hat{y}\|_2^2$$
$$\text{s.t.} \quad \|w\|_1 \leq C .$$

Larger value of λ or smaller value of C result in solutions that have most of their entries equal to or close to 0.

Denoising and Piecewise constant approximations

given signal $x^{ref} \in \mathbb{R}^T$

$x^{ref}(t)$: value of signal at time t, $0 \leq t \leq T, t \in \mathbb{Z}$



Suppose we want to find $y \in \mathbb{R}^T$ s.t. y is a smooth approximation of x^{ref}.

$y(2) \simeq y(1)$
$y(3) \simeq y(2)$ | $y(K) \simeq y(K-1)$

The signal y should have two properties:

i) $y \simeq x^{ref}$ $\quad \|y - x^{ref}\|$ should be small

ii) $[y(k) - y(k-1)]_{k=2}^{T}$ should be small.

$$\min_{y \in \mathbb{R}^T} \|y - x^{ref}\|_2^2 + \lambda \sum_{i=2}^{T} (y(i) - y(i-1))^2$$

In order to achieve sparsity in $[y(k)-y(k-1)]$ vector, we need to use ℓ_1-regularization.

$$\min_{y \in \mathbb{R}^T} \|y - x^{ref}\|_2^2 + \lambda \sum_{i=2}^{T} |y(i) - y(i-1)| .$$

Dual of a Quadratic Program (QP)

Consider the QP

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} x^T H x + c^T x + d \quad , \quad H: \text{symmetric, positive semidef.}$$
$$\text{s.t.} \quad Ax \leq b \quad : \lambda$$

$$\mathcal{L}(x,\lambda) = \frac{1}{2} x^T H x + c^T x + d + \lambda^T (Ax - b)$$
$$= \frac{1}{2} x^T H x + x^T (A^T \lambda + c) + d - \lambda^T b$$

$$d(\lambda) = \inf_{x \in \mathbb{R}^n} \mathcal{L}(x,\lambda) = \begin{cases} -\infty & , \quad A^T\lambda + c \notin R(H) \\ & , \end{cases}$$

$$\nabla_x \mathcal{L}(x,\lambda) = Hx + (A^T\lambda + c) = 0$$
$$\Rightarrow Hx = -(c + A^T\lambda).$$

If $(c + A^T\lambda) \in R(H)$, then $\exists\, z$ s.t. $Hz = -(c + A^T\lambda)$.

$$z = -H^{\dagger}(c + A^T\lambda) .$$

$$d(\lambda) = \frac{1}{2}\left(-H^{\dagger}(c + A^T\lambda)\right)^T H \left(-H^{\dagger}(c + A^T\lambda)\right) - (c + A^T\lambda)^T H^{\dagger} (c + A^T\lambda) + d - \lambda^T b$$

$$= -\frac{1}{2} (c + A^T\lambda)^T H^{\dagger} (c + A^T\lambda) + d - \lambda^T b.$$

dual optimization problem: $\max_{\lambda \geq 0} \; d(\lambda)$

or equivalently $\min_{\lambda \geq 0} \; \underbrace{\frac{1}{2}(c + A^T\lambda)^T H^{\dagger}(c + A^T\lambda) + \lambda^T b - d}_{\text{quadratic in } \lambda}$.

Hence, dual of a QP is another QP.

Quadratically constrained Quadratic programs (QCQP)

$$\min_{x \in \mathbb{R}^n} \; \frac{1}{2} x^T H_0 x + c_0^T x + d_0$$

$$\text{s.t.} \quad \frac{1}{2} x^T H_i x + c_i^T x + d_i \leq 0, \quad i = 1, 2 \cdots m \quad : \lambda_i$$

$$\frac{1}{2} x^T G_i x + f_i^T x + e_i = 0, \quad i = 1, 2 \cdots p \quad : \mu_i$$

The above problem is convex when $H_i \succeq 0, \; i = 0, 1, 2 \cdots m$
$G_i = 0, \; i = 1, 2 \cdots p$

Let us derive its dual when the problem is convex & $H_0 \succ 0$.

$$\mathcal{L}(x, \lambda, \mu) = \frac{1}{2} x^T H_0 x + c_0^T x + d_0 + \sum_{i=1}^{m} \lambda_i \left(\frac{1}{2} x^T H_i x + c_i^T x + d_i\right)$$

$$+ \sum_{j=1}^{p} \mu_i \left(f_i^T x + e_i\right)$$

$$= \frac{1}{2} x^T \underbrace{\left[H_0 + \sum_{i=1}^{m} \lambda_i H_i\right]}_{H(\lambda)} x + x^T \underbrace{\left[c_0 + \sum_{i=1}^{m} \lambda_i c_i + \sum_{j=1}^{p} \mu_i f_i\right]}$$

$$+ \underbrace{\left[d_0 + \sum_{i=1}^{m} \lambda_i d_i + \sum_{j=1}^{p} \mu_i e_i\right]}$$

$$= \frac{1}{2} x^T \underline{H(\lambda)} x + x^T \underline{c(\lambda, \mu)} + \underline{d(\lambda, \mu)}.$$

$\hookrightarrow$ positive definite when $\lambda \geq 0$.

$$\nabla_x \mathcal{L}(x,\lambda,\mu) = H(\lambda)x + c(\lambda,\mu) \Rightarrow x^* = -H(\lambda)^{-1}c(\lambda,\mu)$$

$$d(\lambda,\mu) = -\frac{1}{2}c(\lambda,\mu)^T H(\lambda)^{-1} c(\lambda,\mu) + d(\lambda,\mu)$$

dual:

$$\min_{\lambda,\mu} \quad \frac{1}{2}\underline{c(\lambda,\mu)^T} H(\lambda)^{-1} \underline{c(\lambda,\mu)} - d(\lambda,\mu)$$
$$\text{s.t.} \quad \lambda \geq 0$$

equivalent form:

$$\min_{\lambda,\mu,t} \quad t$$
$$\text{s.t} \quad \frac{1}{2}c(\lambda,\mu)^T H(\lambda)^{-1} c(\lambda,\mu) - d(\lambda,\mu) \leq t$$
$$\lambda \geq 0.$$

$$\begin{bmatrix} [t+d(\lambda,\mu)]2 & c(\lambda,\mu)^T \\ c(\lambda,\mu) & H(\lambda) \end{bmatrix} \succeq 0 \Rightarrow \begin{array}{l} H(\lambda) \succeq 0 \\ 2(t+d(\lambda,\mu)) - c(\lambda,\mu)^T H(\lambda)^{-1} c(\lambda,\mu) \geq 0 \end{array}$$

equivalently:

$$\min_{t,\mu,\lambda} \quad t$$
$$\text{s.t.} \quad \begin{bmatrix} 2(t+\underline{d(\lambda,\mu)}) & \underline{c(\lambda,\mu)^T} \\ \underline{c(\lambda,\mu)} & \underline{H(\lambda)} \end{bmatrix} \succeq 0 :$$
$$\lambda \geq 0.$$

$$\left[F_0 + \sum_{i=1}^{m} \lambda_i F_i + \sum_{j=1}^{k} \mu_j G_j \right]$$

linear matrix inequality, will be discussed in detail subsequently.