Introduction to Machine Learning - CS725
Instructor: Prof. Ganesh Ramakrishnan
Lecture 5a - Least Squares Linear Regression

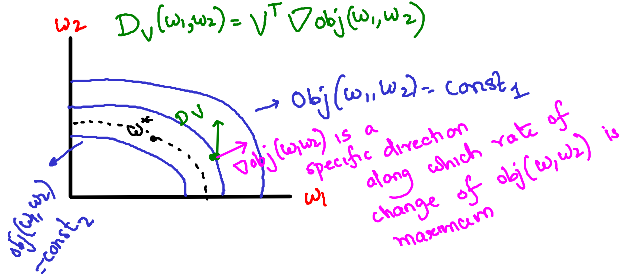
Recall

We recall that the problem was to find $\mathbf{w}$ such that

$$\mathbf{w}^* = \underset{\mathbf{w}}{\text{argmin}} \|\phi\mathbf{w} - \mathbf{y}\|^2 \quad (1)$$

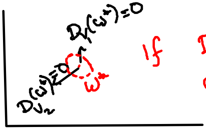
$$= \text{argmin}_{\mathbf{w}}(\mathbf{w}^T\phi^T\phi\mathbf{w} - 2\mathbf{w}^T\phi\mathbf{y} - \mathbf{y}^T\mathbf{y}) \quad (2)$$

$$D_V(\omega_1,\omega_2) = V^T \nabla obj(\omega_1,\omega_2)$$



ω_2
ω_1
$\rightarrow obj(\omega_1,\omega_2) = const_1$
$obj(\omega_1,\omega_2) = const_2$
ω^*
D_V

$\nabla obj(\omega_1,\omega_2)$ is a specific direction along which rate of change of $obj(\omega_1,\omega_2)$ is maximum



If $D_v(w^*)=0$ $\forall$ choices of v & if

$$D_v(w^*) = v^T \nabla f(w^*)$$

Then $\nabla f(w^*)=0$

ie $\frac{\partial f}{\partial w_i}(w^*)=0 \quad \forall i$

Gradient Vector

- Magnitude (euclidean norm) of gradient vector at any point indicates maximum value of directional derivative at that point
- The *gradient vector* of a function f at a point $\mathbf{x}$ is defined as:

$$\nabla f_{\mathbf{x}^*} = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ . \\ . \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} \epsilon \mathbb{R}^n \tag{3}$$

- Thus, at the point of minimum of a differentiable minimization objective (such as least squares for regression), ...

Gradient Vector

- The figure below gives an example of gradient vector

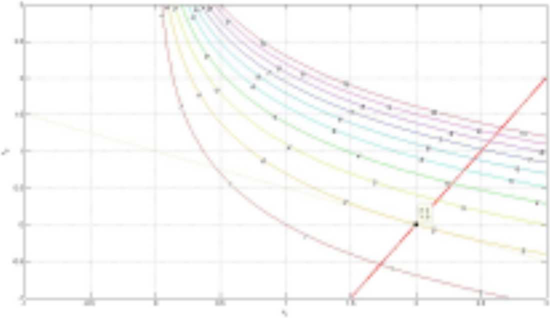


Figure: The level curves along with the gradient vector at (2, 0). Note that the gradient vector is perpendучular to the level curve $x_1 e^{x_2} = 2$ at (2, 0)

Recall

We recall that the problem was to find $\mathbf{w}$ such that

$$\begin{aligned}\mathbf{w}^* &= \underset{\mathbf{w}}{\operatorname{argmin}} \|\phi\mathbf{w} - \mathbf{y}\|^2 && (4)\\ &= \operatorname{argmin}_{\mathbf{w}} (\mathbf{w}^T \phi^T \phi \mathbf{w} - 2\mathbf{w}^T \phi \mathbf{y} - \mathbf{y}^T \mathbf{y}) && (5)\end{aligned}$$

Necessary condition 1

→ ie $\frac{\partial f}{\partial x_i}(x^*)$ exists $\forall i$ [eg: $|x|$ is not differentiable at $x=0$]

- *If* $\nabla f(\mathbf{x}^*)$ *is defined &* $\mathbf{x}^*$ *is local minimum/maximum, then* $\nabla f(\mathbf{x}^*) = 0$ (A necessary condition) `(Cite : Theorem 60)`[1]
- Given that

$\phi = m \times p$

$y = m \times 1$

$w = p \times 1$

$$f(\mathbf{w}) = \underset{\mathbf{w}}{\text{argmin}}(\overset{2}{\cancel{\mathbf{w}^T}}\phi^T\phi\mathbf{w} - 2\cancel{\mathbf{w}^T}\phi\overset{T}{\mathbf{y}} - \cancel{\mathbf{y}^T\mathbf{y}})$$

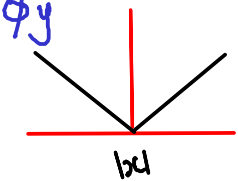
$$\Longrightarrow \ldots\ldots \nabla f(w^*) = 0 = 2\phi^T\phi w - 2\phi^T y$$

- we would have

$$\cancel{2}\phi^T\phi w - \cancel{2}\phi^T y = 0 \ldots\ldots\ldots\ldots$$

$$\Longrightarrow \phi^T\phi w = \phi^T y \ldots\ldots$$

$$\Longrightarrow \ldots\ldots w = (\phi^T\phi)^{-1}\phi^T y$$



[1] **convexopt**.

Necessary condition 1

- *If* $\nabla f(\mathbf{x}^*)$ *is defined &* $\mathbf{x}^*$ *is local minimum/maximum, then* $\nabla f(\mathbf{x}^*) = 0$ (A necessary condition) `(Cite : Theorem 60)`[2]
- Given that

$$\begin{aligned} f(\mathbf{w}) &= \operatorname*{argmin}_{\mathbf{w}}(\mathbf{w}^T\phi^T\phi\mathbf{w} - 2\mathbf{w}^T\phi\mathbf{y} - \mathbf{y}^T\mathbf{y}) && (6) \\ \implies \nabla f(\mathbf{w}) &= 2\phi^T\phi\mathbf{w} - 2\phi^T\mathbf{y} && (7) \end{aligned}$$

- we would have

$$\begin{aligned} \nabla f(\mathbf{w}^*) &= 0 && (8) \\ \implies 2\phi^T\phi\mathbf{w}^* - 2\phi^T\mathbf{y} &= 0 && (9) \\ \implies \mathbf{w}^* &= (\phi^T\phi)^{-1}\phi^T\mathbf{y} && (10) \end{aligned}$$

[2] **convexopt**.

Necessary Condition 2

[Recall from calculus: $\partial^2 f(x) > 0 \Rightarrow \min$]

- *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)
  (Note : Any positive definite matrix is also positive semi-definite)
  `(Cite : Section 3.12 & 3.12.1)`[3]

$$\nabla^2 f(w^*) = \left[\frac{\partial^2 f(w^*)}{\partial w_i \partial w_j}\right]$$

$$\nabla^2 f(w^*) = \nabla(\nabla f(w^*))$$

$$\Longrightarrow$$

$$\nabla f(w) = 2\phi^T\phi w - 2\phi^T y$$

$$\nabla^2 f(w) = 2\phi^T\phi$$

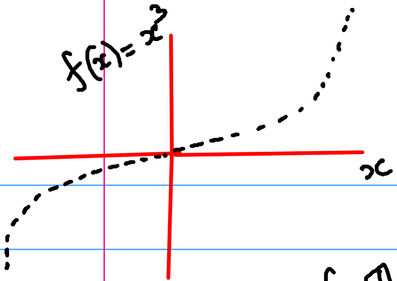
(Note: Hessian $\nabla^2 f$ is in general symmetric... $\frac{\partial^2 f}{\partial w_i \partial w_j} = \frac{\partial^2 f}{\partial w_j \partial w_i}$

- And if ϕ ***has full column rank*** ,

$\phi^T\phi$ is positive definite

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

[3] `cs709/notes/LinearAlgebra.pdf`



$f(0) = 0$

$f'(0) = 0$

$f''(0) = 0$

Though 0 is NOT a point of local min or local max.

Such a pt is called a saddle pt

More generally, x is a saddle pt if

① $\nabla f(x) = 0$ and ② x is neither a local min nor max

Necessary Condition 2

- *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)
  (Note : Any positive definite matrix is also positive semi-definite)
  `(Cite : Section 3.12 & 3.12.1)`[4]

$$\begin{aligned}
\nabla^2 f(\mathbf{w}^*) &= 2\phi^T\phi && (11)\\
\implies \mathbf{x}^T\nabla^2 f(\mathbf{w}^*)\mathbf{x} &= 2\mathbf{x}^T\phi^T\phi\mathbf{x} && (12)\\
&= 2(\phi\mathbf{x})^T\phi\mathbf{x} && (13)\\
&= 2\|\phi\mathbf{x}\|^2 \geq 0 && (14)
\end{aligned}$$

- And if ϕ ***has full column rank*** ,

$$\phi\mathbf{x} = 0 \quad \textit{iff} \quad \mathbf{x} = 0 \qquad (15)$$

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

[4] `cs709/notes/LinearAlgebra.pdf`

Example of linearly correlated features

(when does ϕ not full column rank)

- Example where ϕ doesn't have a full column rank,

$$\phi = \begin{bmatrix} x_1 & x_1^2 & x_1^2 & x_1^3 \\ x_2 & x_2^2 & x_2^2 & x_2^3 \\ \vdots & \vdots & \vdots & \vdots \\ x_n & x_n^2 & x_n^2 & x_n^3 \end{bmatrix} \tag{16}$$

- This is the simplest form of linear correlation of features, and it is not at all desirable.

Some questions

- Based on different inequalities between m and p, what are the cases where the least squares linear regression has (a) no solution (b) one solution and (c) multiple solutions.

ϕ is $m \times p$ ← $\underbrace{A}_{\phi^T\phi}\underset{\rightarrow w}{x} = \underset{\rightarrow \phi^T y}{b}$! r = rank of A & A is $k \times n$

$r = k < n \Rightarrow$ infinit solutions

$r = n < k \Rightarrow$ 1 or 0 solutions

$r = n = k \Rightarrow$ 1 solution

$r < k$ & $r < n \Rightarrow$ 0 or infinit solutions

We discussed:

Necessary condition for local min/max:

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

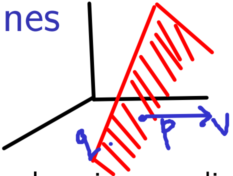
Sufficient condition for local min

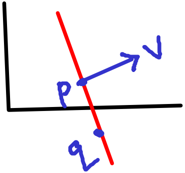
$$\nabla^2 f(w^*) > 0$$

What if $\nabla f(w^*) = 0$ has no closed form soln (such as $w^* = (\phi^T \phi)^{-1} \phi^T y$) eg: when ϕ is not full column rank.. [Gradient descent algo]

What about "global" minimum & conditions for the same? [Convexity]

Hyperplanes



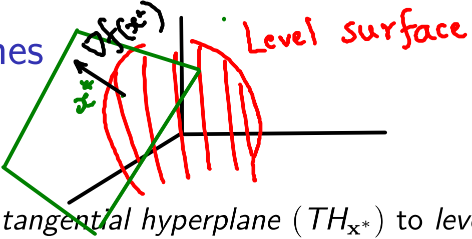


$\mathcal{H}_{v,p}$

- A hyperplane in an n-dimensional Euclidean space is a flat, n-1 dimensional subset of that space that divides the space into two disconnected parts.
- Technically, a hyperplane is a set of points whose direction *w.r.t.* a point $\mathbf{p}$ is orthogonal to a vector $\mathbf{v}$.
- Formally:

$$H_{\mathbf{v},\mathbf{p}} = \left\{ \mathbf{q} \mid (\mathbf{p} - \mathbf{q})^{\mathbf{T}} \mathbf{v} = \mathbf{0} \right\} \tag{17}$$

Tangential Hyperplanes



There are two definitions of *tangential hyperplane* $(TH_{\mathbf{x}^*})$ to *level surface* $(L_{f(\mathbf{x}^*)}(f))$ of f at $\mathbf{x}^*$:

- Plane consisting of all tangent lines at $\mathbf{x}^*$ to any parametric curve $c(t)$ on level surface.
- Plane orthogonal to the gradient vector at $\mathbf{x}^*$.

$$\mathcal{H}_{x^*, \nabla f(x^*)} = TH_{\mathbf{x}^*} = \left\{ \mathbf{p} \mid (\mathbf{p} - \mathbf{x}^*)^{\mathbf{T}} \nabla \mathbf{f}(\mathbf{x}^*) = \mathbf{0} \right\} \qquad (18)$$

Idea of descent algos:- $\Delta x^* = \operatorname{argmin} -\nabla^T f(x) \Delta x$ s.t $\Omega(\Delta x) \leq \theta$

$\Delta x^* =$ See on next page $\Leftarrow \Omega(\Delta x) = \max(|\Delta x|)$

$\Omega(\Delta x) = \|\Delta x\|_2^2 \Rightarrow \Delta x^* = -\nabla f(x)$

Q: Consider $\max -\nabla^T f(x)\,\Delta x$

$\text{s.t.} \quad \max(|\Delta x|) \le \Theta$

Why should $\Delta x = -\nabla f(x)$ not be soln?

$\nabla^T f(x)\,\Delta x = \dfrac{\Theta\,\|\nabla f(x)\|^2}{\max(|\nabla f(x)|)}$ $\left(\dfrac{\max(|\nabla f(x)|)}{\Theta}\right)$ $\Omega(\Delta x) = \Theta$

Ans: Consider

$$\Delta x' = \begin{bmatrix} \Theta\,\text{sgn}[(\nabla f(x))_1] \\ \Theta\,\text{sgn}[(\nabla f(x))_2] \\ \Theta\,\text{sign}[(\nabla f(x))_i] \end{bmatrix}$$

$\Omega(\Delta x') = \Theta$

$\nabla^T f(x)\,\Delta x' = \Theta \sum_{i=1}^{\ell} |(\nabla f(x))_i|$

CLAIM: $-\nabla^T f(x)\,\Delta x' < -\nabla^T f(x)\,\Delta x$ $\Big\}$ Though $\Omega(\Delta x) = \Omega(\Delta x')$

Gradient Descent Algorithm

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

$$\underline{\Delta \mathbf{w}^{(\mathbf{k})} = -\nabla\varepsilon(\mathbf{w}^{(\mathbf{k})})} \qquad \textbf{from equation } (\mathbf{10})$$

$w^{(0)}$ = random vector

Hence,

$$\mathbf{w}^{(\mathbf{k+1})} = \mathbf{w}^{(\mathbf{k})} + \underset{\text{step length}}{\mathbf{2t}^{(\mathbf{k})}}\underbrace{(\phi^{\mathbf{T}}\mathbf{y} - \phi^{\mathbf{T}}\phi\mathbf{w}^{(\mathbf{k})})}_{-\nabla\varepsilon(w^{(k)})} \tag{19}$$

Gradient Descent Algorithm

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

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

magnitude of gradient should nearly vanish

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 $\mathbf{x}$ should move along that direction.
- In general,

$$t^{(k)} = \underset{t}{\operatorname{argmin}} f\left(\mathbf{w}^{(\mathbf{k+1})}\right) \tag{20}$$

- Thus,

one dimensional opt problem

$$t^{(k)} = \underset{t}{\operatorname{argmin}} f\left(\mathbf{w}^{(\mathbf{k})} + \mathbf{2t}\left(\phi^{\mathbf{T}}\mathbf{y} - \phi^{\mathbf{T}}\phi\mathbf{w}^{(\mathbf{k})}\right)\right) \tag{21}$$

Example of Gradient Descent Algorithm

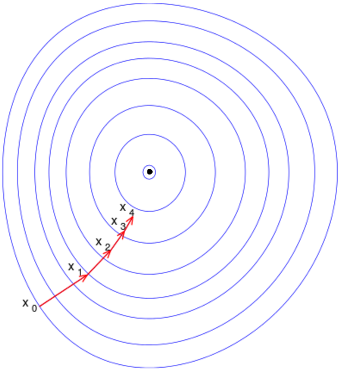


Figure: A red arrow originating at a point shows the direction of the negative gradient at that point. Note that the (negative) gradient at a point is orthogonal to the level curve going through that point. We see that gradient descent leads us to the bottom of the bowl, that is, to the point where the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