Convex Optimization, Constrained Optimization and Regression

Instructor: Prof. Ganesh Ramakrishnan

Agenda

So far, conditions such as: $\nabla f(w^*) = 0$
$\underline{\underline{\text{OR}}}$ $\nabla^2 f(w^*) > 0$ were conditions for "local" min/max

- Definition of Convex Sets and Functions
- Example of Convex Set
- Example of Convex Function
- Theorem related to Convex Functions
- Overfitting
- Convex Optimization Problems
- Next Lecture

Global min/max

Convex combinations lie within the same dmn

Definition of Convex Sets and Convex Functions

Definition of convex sets and convex functions (`Cite : Definition 32 and 35`)[1]

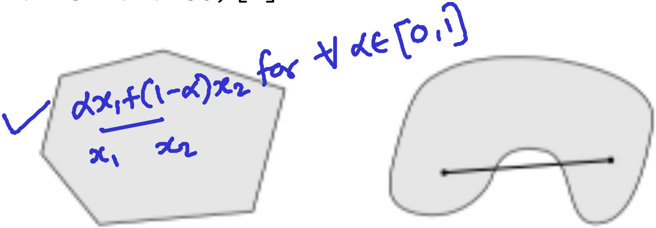


Figure: Examples of a convex set (a) and a non-convex set (b) Cite: http://cs229.stanford.edu/section/cs229-cvxopt.pdf

A set C is convex if, for any x,y $\in$ C and $\theta \in \Re$ and $0 \leq \theta \leq 1$,

$$\theta x + (1-\theta)\cancel{x}y \in C \tag{1}$$

Example of a Convex Set

$$\mathcal{H}_{p,v} = \{q \mid (p-q)^T v = 0\}$$

Verify by: $\underbrace{q_1 \in H_{p,v}}_{(p-q_1)^T v = 0}$ $\underbrace{q_2 \in H_{p,v}}_{(p-q_2)^T v = 0} \Rightarrow \theta q_1 + (1-\theta) q_2 \in H_{p,v}$

$\Rightarrow$ - - - - .

To prove : *Verify that a hyperplane is a convex set.*

Proof

- A Hyperplane $\mathcal{H}$ is defined as $\{\mathbf{x}|\mathbf{a}^T\mathbf{x} = b, \mathbf{a} \neq \mathbf{0}\}$
- Let $\mathbf{x}$ and $\mathbf{y}$ be vectors that belong to the hyperplane
- Since they belong to the hyperplane, $\mathbf{a}^T\mathbf{x} = b$ *and* $\mathbf{a}^T\mathbf{y} = b$
- In order to prove the convexity of the set we must show that :

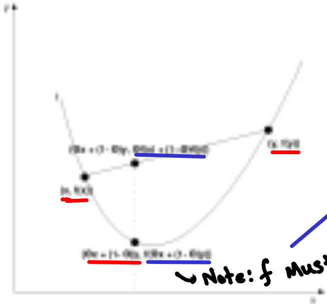
$$\theta\mathbf{x} + (1-\theta)\mathbf{y} \in \mathcal{H}, \; \textit{where } \theta \in [0,1] \tag{2}$$

- In particular, it will belong to the hyperplane if it's true that :

$$\begin{aligned} \mathbf{a}^T(\theta\mathbf{x} + (1-\theta)\mathbf{y}) &= b & (3)\\ \implies \mathbf{a}^T\theta\mathbf{x} + \mathbf{a}^T(1-\theta)\mathbf{y} &= b & (4)\\ \implies \theta\mathbf{a}^T\mathbf{x} + (1-\theta)\mathbf{a}^T\mathbf{y} &= b & (5) \end{aligned}$$

- And, we also have $\mathbf{a}^T\mathbf{x} = b$ *and* $\mathbf{a}^T\mathbf{y} = b$. Hence $\theta b + (1-\theta)b = b$. [Hence Proved] So a hyperplane is a convex set.

Definition of Convex Sets and Convex Functions



Note: f must be defined at each convex comb $\theta x + (1-\theta) y$ ⇒ dmn must be convex!

Figure: A sample convex function

$$\therefore f(\theta \mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y}) \tag{6}$$

$\forall\, x, y \in dmn(f)$

Assuming dmn(f) is convex

Example of a Convex Function

Q. *Is* $||\phi\mathbf{w} - \mathbf{y}||^2$ *convex?* (in w)

A.

- To check this, we have `(Cite : Theorem 75)`[1]. Is this practical? : $f(w') \geq f(w) + \nabla^T f(w)(w' - w) \ \forall w, w'$
- Instead, we would use `(Cite : Theorem 79)`[2] to check for the convexity of our function : $\nabla^2 f(w) \succeq 0$ (p.s.d) $\forall w$
- So the condition that has our focus is -

$$\nabla^2 f(\mathbf{w}^*) \ \textit{is positive semi} - \textit{definite}, \ \textit{if} \ \forall \mathbf{x} \neq 0, \ \mathbf{x}^T \nabla^2 f(\mathbf{w}^*)\mathbf{x} \geq 0 \tag{7}$$

- We have, is always p.s.d even if ϕ is NOT full colmn rank!

$$\nabla^2 f(\mathbf{w}) = 2\phi^T\phi \tag{8}$$

$\rightarrow$ independent of w

- So, $||\phi\mathbf{w} - \mathbf{y}||^2$ is convex, since the domain for $\mathbf{w}$ is $\mathbb{R}^n$ and is convex

[1] `cs709/notes/BasicsOfConvexOptimization.pdf`

[2] `cs709/notes/BasicsOfConvexOptimization.pdf`

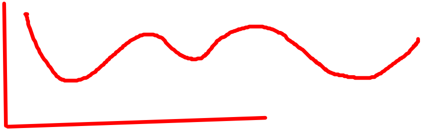
Strict Convexity

Eg: $f(x) = a^T x + b$ is convex but NOT strictly convex

Q. *When is* $f(\mathbf{x})$ *(strictly) convex?*
A1. Iff $f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \leq \; (<) \; \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y})$ for all $\theta \in [0,1]$ and for all $\mathbf{x}, \mathbf{y} \in dmn(f)$
A2. OR Iff $\nabla^2 f(\mathbf{x})$ is positive semi-definite (definite) for all $\mathbf{x} \in dmn(f)$

Q: When is $\|\phi w - y\|^2$ strictly convex?
Ans: When ϕ is full column rank so that $\phi^T\phi$ is positive definite

Strict Convexity



Q. *When is $f(\mathbf{x})$ (strictly) convex?*
A1. Iff $f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \leq \ (<)\ \theta f(\mathbf{x}) + (1-\theta)f(\mathbf{y})$ for all $\theta \in [0,1]$ and for all $\mathbf{x}, \mathbf{y} \in dmn(f)$
A2. OR Iff $\nabla^2 f(\mathbf{x})$ is positive semi-definite (definite) for all $\mathbf{x} \in dmn(f)$
Q. *Is $\|\phi\mathbf{w} - \mathbf{y}\|^2$ strictly convex?*
A. Iff ϕ has full column rank.

To prove: *If a function is convex, any point of local minima $\equiv$ point of global minima*

Does not hold for maxima

Proof - `(Cite : Theorem 69)`[3]

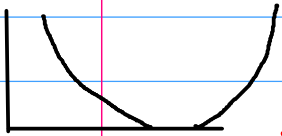
Thus: $w^* = (\phi^T\phi)^{-1}\phi^T y$ is unique global minimizer of $\|\phi w - y\|_2^2$

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

If a function f is strictly convex, it will have a unique global minimum!

Eg: $f(x) = b$ ... is convex but not strictly convex

It has global min = local min = all pts in $dom(f)$



Not strictly convex

Theorem

To prove : *If a function is strictly convex, it has a unique point of global minima*
Proof - `(Cite : Theorem 70)`[4]
Since $\|\phi\mathbf{w} - \mathbf{y}\|^2$ is strictly convex for linearly independent ϕ,

$$\nabla f(\mathbf{w}^*) = 0 \textit{ for } \mathbf{w}^* = (\phi^T\phi)^{-1}\phi^T\mathbf{y} \tag{9}$$

Thus, $\mathbf{w}^*$ is a point of global minimum. One can also find a solution to $(\phi^T\phi\mathbf{w} = \phi^T\mathbf{y})$ by Gauss elimination.

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

Redundant ϕ and Overfitting

- What do you expect in experiments if ϕ had redundancy?

$\phi_2 \equiv \phi_3$

$\phi_3 = \phi_1 + \phi_2$

Linear redundancy

$\underbrace{\phi_2 = \phi_1^2 \quad \phi_4 = \phi_1^2 + \phi_2^3}_{\text{non-linear redundancy}}$

Example of linearly correlated features

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

same

$$\phi = \begin{bmatrix} x_1 & x_1^2 & x_1^2 & x_1^3 \\ x_2 & x_2^2 & x_2^2 & x_2^3 \\ \vdots & \vdots & \vdots & \vdots \\ x_n & x_n^2 & x_n^2 & x_n^3 \end{bmatrix} \tag{10}$$

- This is the simplest form of linear correlation of features.

Redundant ϕ and Overfitting

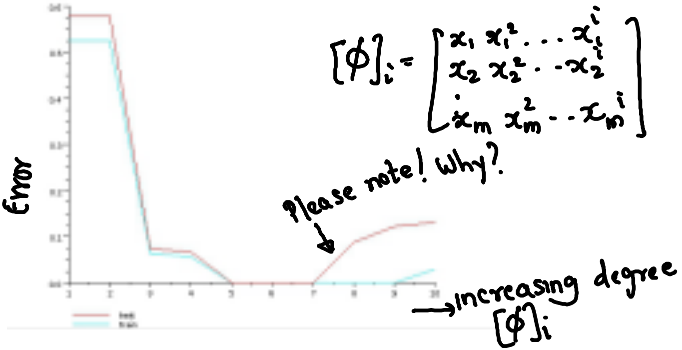


Figure: train-RMS and test-RMS values vs t(degree of polynomial) graph

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

Homework:
Explain why the error on the train data reduces as the degree increases until 7. Why does the error on the test data also decrease until degree of 7?

Now explain why the train continues to remain low even beyond degree of 7 whereas the test data starts increasing now.