WEEK 9: REVISION

FINAL EXAM

CONTENTS

1. Properties of Convex Functions
2. Applications of Optimization in Machine Learning
3. Revisiting Constrained Optimization
4. Relation between Primal and Dual Problem, KKT Conditions
5. KKT conditions continued

1. PROPERTIES OF CONVEX FUNCTIONS

Necessary and sufficient conditions for optimality of convex functions

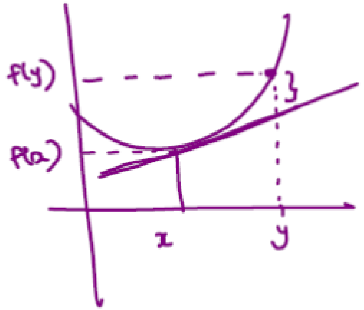
Goal : $\min_{x} f(x)$

Theorem

Let f be a differentiable and convex function from $\mathbb{R}^d \to \mathbb{R}$, $x^* \in \mathbb{R}^d$ is a global minimum of f **if and only if** $\nabla f(x^*) = 0$.

If $\exists\, x^*$ s.t $\nabla f(x^*) = 0$ $\Rightarrow$ x^* is a global minima.

By definition of convexity



$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \forall\, x, y.$$

$$\Rightarrow \quad f(y) \geq f(x^*) + \underbrace{\nabla f(x^*)^T (y - x)}_{0} \quad \forall y$$

$(\because \nabla f(x^*) = 0$ by assumption$)$

$$\Rightarrow \quad \boxed{f(y) \geq f(x^*) \quad \forall y} \quad \Rightarrow \quad x^* \text{ is a global minima!}$$

1. PROPERTIES OF CONVEX FUNCTIONS

If $f : \mathbb{R}^d \to \mathbb{R}, g : \mathbb{R}^d \to \mathbb{R}$ are both convex functions, then $f(x) + g(x)$ is a convex function

Proof: Fix ($\lambda \in [0,1]$)

$$h(\lambda x + (1-\lambda)y) = \underbrace{f(\lambda x + (1-\lambda)y)} + \underbrace{g(\lambda x + (1-\lambda)y)}$$

[by defn of h]

$$\leq \lambda f(x) + (1-\lambda) f(y) + \lambda g(x) + (1-\lambda) g(y)$$

[by convexity of f and g]

$$= \lambda (f(x) + g(x)) + (1-\lambda)(f(y) + g(y))$$

$$= \lambda h(x) + (1-\lambda) h(y)$$

$\boxed{\text{Sums of convex functions is convex}}$

1. PROPERTIES OF CONVEX FUNCTIONS

Let $f : \mathbb{R} \to \mathbb{R}$ is a convex and non-decreasing function and $g : \mathbb{R}^d \to \mathbb{R}$ be a convex function, then their composition $h = f(g(x))$ is also a convex function.

Fix $\lambda \in [0,1]$; $\boxed{h(\lambda x + (1-\lambda) y)}$

$$= f(\underbrace{g(\lambda x + (1-\lambda) y)}_{z}) \quad z \leq \overbrace{\lambda g(x) + (1-\lambda) g(y)}^{z'}$$

$$\leq f(\lambda \underbrace{g(x)}_{x'} + (1-\lambda) \underbrace{g(y)}_{y'})$$

[by convexity of g and by non-decreasing property of f]

$$\leq \lambda f(g(x)) + (1-\lambda) f(g(y))$$

$$= \lambda h(x) + (1-\lambda) h(y).$$

Composition of convex with convex + non-decreasing is convex.

1. PROPERTIES OF CONVEX FUNCTIONS

Let $f:\mathbb{R}\to\mathbb{R}$ is a convex function and $g:\mathbb{R}^d\to\mathbb{R}$ be a linear function, then their composition $h=f(g(x))$ is also a convex function.

Proof: Fix $\lambda\in[0,1]$. $h(\lambda x+(1-\lambda)y)$

$$= f\big(g(\lambda x+(1-\lambda)y)\big)$$

$$= f\big(\lambda g(x)+(1-\lambda)g(y)\big) \quad [\text{by linearity of } g]$$

$$\boxed{\leq} \ \lambda f(g(x))+(1-\lambda)f(g(y)) = \lambda h(x)+(1-\lambda)h(y)$$

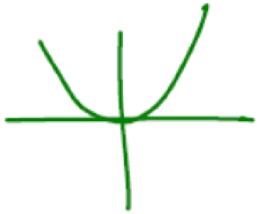
[Convexity of f]

1. PROPERTIES OF CONVEX FUNCTIONS

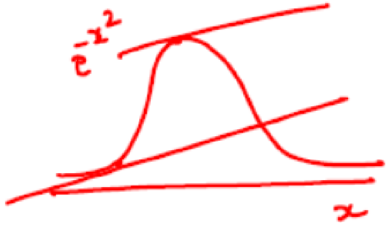
In general, if f and g are both convex functions, then $h = fog$ may not be convex function.

$g(x) = x^2 \rightarrow$ Convex

$f(x) = e^{\boxed{-x}} \rightarrow$ Convex.



$f \circ g(x) = e^{-x^2}$. is not Convex.



Note: g is concave if and only if $f = -g$ is convex.

2. APPLICATIONS OF OPTIMIZATION IN ML

Linear Regression:

Training data $\rightarrow$ $X_1, X_2, \ldots, X_n$ with corresponding outputs $y_1, y_2, \ldots, y_n$, where $X_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, $\forall i$.

Performance measure : Sum of squares error.

$$\min_{w \in \mathbb{R}^d} \underbrace{\sum_{i=1}^{n} (\underbrace{w^T x_i - y_i}_{\text{error of } x_i \text{ made by } w})^2}_{f(w)}$$

Specific Goal of linear regression

$$\boxed{\min_{w \in \mathbb{R}^d} \frac{1}{2} \sum_{i=1}^{n} (w^T x_i - y_i)^2}$$

Gradient of the sum of squares error

$$\boxed{\nabla f(w) = (x^T x) w - x^T y}$$

Analytical or closed form solution of coefficients w^* of a linear regression model

$$w^* = (X^T X)^{-1} X^T y$$

2. APPLICATIONS OF OPTIMIZATION IN ML

→ GD will reach GM if GM ⇒ LM

In linear regression, the gradient descent approach avoids the inverse computation by iteratively updating the weights.

$$w^{t+1} = w^t - \eta_t \nabla f(w^t))$$

$$w^{t+1} = w^t - \eta_t((X^T X)w^t - X^T y)$$

10,000 x 10,000

Stochastic gradient descent:

- Computes approximation of gradient to make gradient computation faster (because in GD $X^T X$ will use entire dataset).
- Samples a small set of data points at random for every iteration to compute the gradient.

$\frac{1}{T}\sum_{t=1}^{T} w_t \rightarrow w^*$

3. CONSTRAINED OPTIMIZATION

Consider the constrained optimization problem as follows:

$$\min_x f(x)$$
$$\text{subject to } h(x) \leq 0$$

Lagrangian function:

$$\underset{\substack{\uparrow \\ \text{vector}}}{\mathcal{L}(x}, \underset{\substack{\uparrow \\ \text{Scalar}}}{\lambda)} = f(x) + \lambda h(x)$$

$$\boxed{\begin{array}{l} \min_x f(x) \\ \text{s.t } h(x) \leq 0 \end{array}} \equiv \boxed{\min_x \left[\max_{\lambda \geq 0} \mathcal{L}(x, \lambda) \right]}$$

primal

4. RELATION BETWEEN PRIMAL AND DUAL PROBLEM

x^* → $\min_{x} \left[\max_{\lambda \geq 0} \; L(x,\lambda) \right]$

Primal

$\max_{\lambda \geq 0} \; \min_{x} \; L(x,\lambda)$ ← λ^*

DUAL

Weak Duality	Strong Duality
$g(\lambda^*) \leq f(x^*)$	If f and g are convex functions. $g(\lambda^*) = f(x^*)$

5. KARUSH-KUHN-TUCKER CONDITIONS

Consider the optimization problem with multiple equality and inequality constraints as follows:

$$\min_x f(x)$$
subject to
$$h_i(x) \leq 0, \forall i = 1, \dots, m$$
$$l_j(x) = 0, \forall j = 1, \dots, n$$

The Lagrangian function is expressed as follows:

$$L(x, u, v) = f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{n} v_j l_j(x)$$

Karush-Kuhn-Tucker Conditions:

Stationarity $\boxed{\nabla f(x) + \sum_{i=1}^{n} u_i \nabla h(x) + \sum_{j=1}^{m} v_j \nabla l(x) = 0}$

Complementary slackness $\boxed{u_i h_i = 0}$ $\forall i$

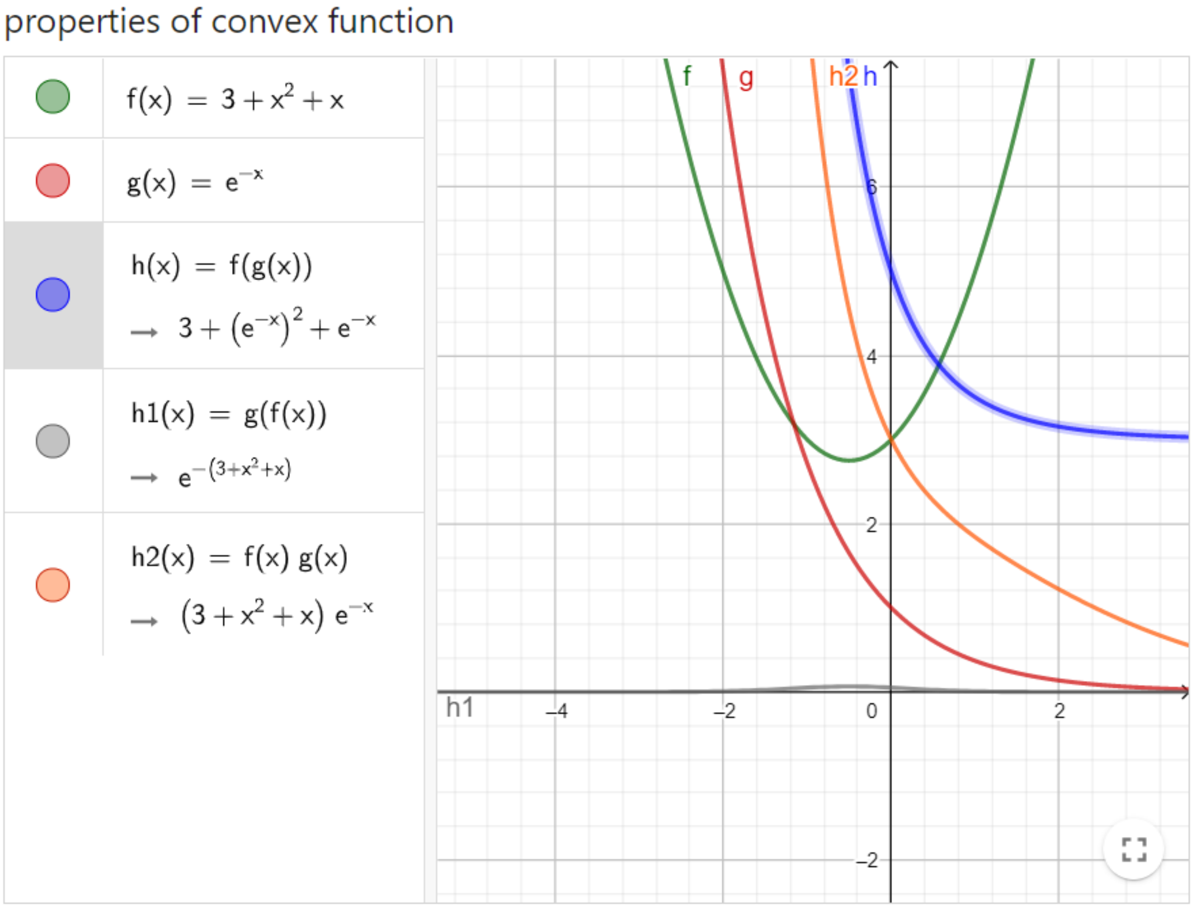
Primal feasibility $\boxed{h_i(\mathrm{x}) \leq 0}$ $\forall i$

Dual feasibility $\boxed{u_i \geq 0}$ $\forall i$

SOME SOLVED PROBLEMS

Properties of convex functions

https://www.geogebra.org/m/esqcd4he



Given below is a set of data points and their labels.

f_1 f_2

X	y
[1,0]	1.5
[2,1]	2.9
[3,2]	3.4
[4,2]	3.8
[5,3]	5.3

How to find the optimal w^* using the analytical method?

Let us use Gradient descent optimization.

optimal $w^* = (X^T X)^{-1}(X^T y)$

$$X \begin{bmatrix} 1 & 0 \\ 2 & 1 \\ 3 & 2 \\ 4 & 2 \\ 5 & 3 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 55 & 31 \\ 31 & 18 \end{bmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} 59.2 \\ 33.2 \end{pmatrix}$$

$$X^T y = \begin{bmatrix} 59.2 \\ 33.2 \end{bmatrix}$$

$$w^* = \begin{bmatrix} 1.255 \\ -0.317 \end{bmatrix}$$

$$\text{Given } w^1 = \begin{bmatrix} 0.1 \\ 0.1 \end{bmatrix}$$

$$\text{Gradient } \boxed{\nabla f(w) = (X^T X)w - X^T y}$$

$$\nabla f(w) = \begin{bmatrix} -50.6 \\ -28.3 \end{bmatrix}$$

$$\text{update equation: } w^2 = w^1 - \eta_t \nabla f(w^1)$$

$$w^2 = \begin{bmatrix} 0.1 \\ 0.1 \end{bmatrix} - 0.1 \begin{bmatrix} -50.6 \\ -28.3 \end{bmatrix}$$

$$\boxed{w^2 = \begin{bmatrix} 5.16 \\ 2.93 \end{bmatrix}}$$

$$\textit{minimize} \quad 3x_1 + x_2 \rightarrow f \qquad \nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{pmatrix}$$

subject to

$$u_1 \qquad x_1 - x_2 + 4 \leq 0 \rightarrow g_1 \qquad \nabla g_1$$

$$u_2 \qquad -3x_1 + 2x_2 + 10 \leq 0 \rightarrow g_2 \qquad \nabla g_2$$

Stationarity conditions

$$3 + u_1 - 3u_2 = 0 \quad \text{— } 1$$

$$1 - u_1 + 2u_2 = 0 \quad \text{— } 2$$

Complementary slackness conditions

$$u_1(x_1 - x_2 + 4) = 0 \quad \text{— } 3$$

$$u_2(-3x_1 + 2x_2 + 10) = 0 \quad \text{— } 4$$

Primal feasibility conditions

$$x_1 - x_2 + 4 \leq 0$$

$$-3x_1 + 2x_2 + 10 \leq 0$$

Dual feasibility conditions

$$u_1, u_2 \geq 0$$

	$u_2 = 0$	$u_2 \neq 0$
$u_1 = 0$	x	x
$u_1 \neq 0$	x	✓

Case (i) $u_1 = 0$; $u_2 = 0$

$3 \neq 0$

$1 \neq 0$

Case (ii) $u_1 = 0$, $u_2 \neq 0$

$3 - 3u_2 = 0 \qquad u_2 = 1$ ✓

$1 + 2u_2 = 0$

$-3x_1 + 2x_2 + 10 = 0$

Case (iii) $u_1 \neq 0$, $u_2 = 0$

$3 + u_1 = 0$

$1 - u_1 = 0 \qquad u_1 = 1$ ✓

$x_1 - x_2 + 4 = 0$

$$minimize \quad 3x_1 + x_2$$

$$subject\ to$$

$$x_1 - x_2 + 4 \leq 0$$

$$-3x_1 + 2x_2 + 10 \leq 0$$

Stationarity conditions

$$3 + \cancel{u_1} - 3u_2 = 0 \quad (1)$$

$$1 - \cancel{u_1} + 2\underbrace{u_2}_{=4} = 0 \quad (2)$$

Complementary slackness conditions

$$\overset{\neq 0}{u_1}(x_1 - x_2 + 4) = 0 \quad (3)$$

$$\overset{\neq 0}{u_2}(-3x_1 + 2x_2 + 10) = 0 \quad (4)$$

Primal feasibility conditions

$$x_1 - x_2 + 4 \leq 0$$

$$-3x_1 + 2x_2 + 10 \leq 0$$

Dual feasibility conditions

$$u_1, u_2 \geq 0$$

$u_1 \neq 0 \quad u_2 \neq 0$

$4 - u_2 = 0 \quad \therefore \boxed{u_2 = 4, \quad u_1 = 9}$

$x_1 - x_2 + 4 = 0 \Rightarrow 2x_1 - \cancel{2x_2} + 8 = 0$

$-3x_1 + \cancel{2x_2} + 10 = 0$

$-x_1 + 18 = 0 \quad \boxed{\therefore x_1 = 18 \quad \therefore x_2 = 22}$

minimum $f(x_1, x_2) = 76$


Any
Questions


thank you