1

(a) Starting from the definition of a convex function

$$f(ax + (1-a)y) \le a\,f(x) + (1-a)\,f(y) \tag{1}$$

Let $a = \epsilon/(x-y)$ and rearrange the inequality to give

$$(x-y)\left(\frac{f(y+\epsilon) - f(y)}{\epsilon}\right)$$

on the left-hand side. Taking the limit $\epsilon \to 0$ show that the function $f(x)$ lies above the tangent line $t(x) = f(y) + (x-y)\,f'(y)$ going through the point y. [4 marks]

Rearranging the Equation 1

$$f(y + a(x-y)) \le f(y) + a\,(f(x) - f(y)).$$

Or

$$\frac{f(y + a(x-y)) - f(y)}{a} \le f(x) - f(y).$$

Letting $a = \epsilon/(x-y)$ then

$$(x-y)\,\frac{f(y+\epsilon) - f(y)}{\epsilon} \le f(x) - f(y)$$

Taking the limit $\epsilon \to 0$ then using

$$\lim_{\epsilon\to 0} \frac{f(y+\epsilon) - f(y)}{\epsilon} = f'(y)$$

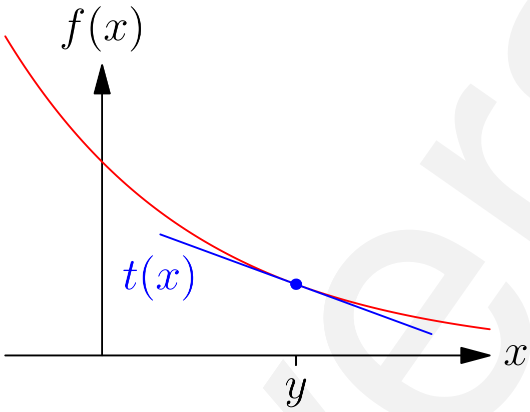
So that

$$(x-y)\,f'(y) \le f(x) - f(y)$$

or

$$f(x) \ge f(y) + (x-y)\,f'(y) = t(x).$$

(b) Sketch the tangent line, $t(x)$, at the point y in the graph shown below. [1 mark]



(c) Starting from the inequality for a convex function

$$f(x) \geq f(y) + (x - y)\, f'(y) \tag{2}$$

consider the case $y = x + \epsilon$, then by Taylor expanding $f(x + \epsilon)$ and $f'(x + \epsilon)$ around x and keeping all terms up to order ϵ^2 show that for a convex function $f''(x) \geq 0$. [4 marks]

We use the expansions

$$f(x + \epsilon) = f(x) + \epsilon f'(x) + \frac{\epsilon^2}{2} f''(x) + O(\epsilon^3)$$
$$f'(x + \epsilon) = f'(x) + \epsilon f''(x) + O(\epsilon^2)$$

Using $y = x + \epsilon$ and substituting into the Equation (2)

$$f(x) \geq f(x + \epsilon) - \epsilon f'(x + \epsilon)$$
$$f(x) \geq f(x) + \epsilon f'(x) + \frac{\epsilon^2}{2} f''(x) + O(\epsilon^3) - \epsilon \left(f'(x) + \epsilon f''(x) + O(\epsilon^2) \right)$$

Or subtraction $f(x)$ on both sides

$$0 \geq -\frac{\epsilon^2}{2} f''(x) + O(\epsilon^3)$$

Since this has to be true for all $\epsilon > 0$ this requires $f''(x) \geq 0$.

(d) Prove that $-\log(x)$ is convex-up for $x > 0$. [1 mark]

$$\frac{\mathrm{d}^2(-\log(x))}{\mathrm{d}x^2} = \frac{1}{x^2} \geq 0$$

End of question 1

2

(a) If $\|\boldsymbol{x}\|$ is a proper norm use the triangular inequality ($\|\boldsymbol{x}+\boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$), linearity of a norm ($\|a\boldsymbol{x}\| = a\|\boldsymbol{x}\|$) and the definition of convexity, to show that the norm is convex. [5 marks]

For any vectors, $\boldsymbol{x}$, and $\boldsymbol{y}$ and any scale $a \in [0,1]$ then

$$\|a\boldsymbol{x} + (1-a)\boldsymbol{y}\| \leq \|a\boldsymbol{x}\| + \|(1-a)\boldsymbol{y}\| = a\|\boldsymbol{x}\| + (1-a)\|\boldsymbol{y}\|$$

where the first inequality follows from the triangular inequality and the second equality from the linearity of the norm. However,

$$\|a\boldsymbol{x} + (1-a)\boldsymbol{y}\| \leq a\|\boldsymbol{x}\| + (1-a)\|\boldsymbol{y}\|$$

is the defining equation of convexity.

(b) Consider a classification problem where $\hat{f}_c(\boldsymbol{x}|\boldsymbol{\theta})$ is the probability that a learning machine with parameters $\boldsymbol{\theta}$ predicts that input $\boldsymbol{x}$ belongs to class $c \in \mathcal{C}$. Assume the training is stochastic so the probability of obtaining parameters $\boldsymbol{\theta}$ is $\rho(\boldsymbol{\theta})$. Let $\hat{m}_c(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\theta}}\left[\hat{f}_c(\boldsymbol{x}|\boldsymbol{\theta})\right]$ be the output of the mean machine for class c. Assuming that for a data point $(\boldsymbol{x}, y)$, where y is a class label, we use a cross entropy loss

$$L(\boldsymbol{x}, y, \boldsymbol{\theta}) = -\sum_{c \in \mathcal{C}} [\![y = c]\!] \log\left(\hat{f}_c(\boldsymbol{x}|\boldsymbol{\theta})\right),$$

show that the expected loss over inputs and parameters can be written as the expected loss of the mean machine plus a second loss. Use Jensen's inequality ($\mathbb{E}\left[\log(X)\right] \leq \log\left(\mathbb{E}\left[X\right]\right)$) to show the second term is positive. [5 marks]

$$\begin{aligned}
\bar{L} &= \mathbb{E}_{(\boldsymbol{x},y)}\left[\mathbb{E}_{\boldsymbol{\theta}}\left[-\sum_{c \in \mathcal{C}} [\![y = c]\!] \log\left(\hat{f}_c(\boldsymbol{x}|\boldsymbol{\theta})\right)\right]\right] \\
&= -\mathbb{E}_{(\boldsymbol{x},y)}\left[\sum_{c \in \mathcal{C}} [\![y = c]\!] \log(\hat{m}_c(\boldsymbol{x}))\right] - \mathbb{E}_{(\boldsymbol{x},y)}\left[\mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{c \in \mathcal{C}} [\![y = c]\!] \log\left(\frac{\hat{f}_c(\boldsymbol{x}|\boldsymbol{\theta})}{\hat{m}_c(\boldsymbol{x})}\right)\right]\right]
\end{aligned}$$

The first terms acts like a bias. The second (variance-like) term is

$$-\mathbb{E}_{(\boldsymbol{x},y)}\left[\sum_{c \in \mathcal{C}} [\![y = c]\!] \mathbb{E}_{\boldsymbol{\theta}}\left[\log\left(\hat{f}_c(\boldsymbol{x}|\boldsymbol{\theta})\right)\right]\right] + \mathbb{E}_{(\boldsymbol{x},y)}\left[\sum_{c \in \mathcal{C}} [\![y = c]\!] \log(\hat{m}_c(\boldsymbol{x}))\right]$$

But using Jensen's inequality
$\mathbb{E}_{\boldsymbol{\theta}}\left[\log\left(\hat{f}_c(\boldsymbol{x}|\boldsymbol{\theta})\right)\right] \leq \log\left(\mathbb{E}_{\boldsymbol{\theta}}\left[\hat{f}_c(\boldsymbol{x}|\boldsymbol{\theta})\right]\right) = \log(\hat{m}_c(\boldsymbol{x}))$. Thus this second term is positive.

End of question 2

END OF PAPER