NONE

ADRIAN KUKLA HW2 904056948

Qs 2.1 Activation Function

a) $g(-x) = \frac{-x}{1+|-x|} = \frac{-x}{1+|x|}$

$-g(x) = \frac{-x}{1+|x|} = g(-x)$ Hence $g(x)$ is centered at zero.

b) i) For $x>0$, $|x| = x$

$g(x) = \frac{x}{1+x}$ ~~$g'(x) = 1+x$~~ ~~$\frac{d}{dx}\left[\frac{f(x)}{g(x)}\right] = \frac{g(x)f'(x) - f(x)g'(x)}{(g(x))^2}$~~

$g'(x) = \frac{(1+x)\times 1 - x\times 1}{(1+x)^2} = \frac{1}{(1+x)^2}$

For $x<0$, $|x| = -x \Rightarrow g(x) = \frac{x}{1-x}$

$g'(x) = \frac{(1-x)\cdot 1 - x(-1)}{(1-x)^2} = \frac{1}{(1-x)^2}$

$$g'(x) = \begin{cases} \frac{1}{(1+x)^2}, & \text{for } x>0 \\ \frac{1}{(1-x)^2}, & \text{for } x<0 \end{cases}$$

ii) $\lim_{x\to 0^+} g'(x) = \lim_{x\to 0^+} \frac{1}{(1+x)^2} = 1$

$\lim_{x\to 0^-} g'(x) = \lim_{x\to 0^-} \frac{1}{(1-x)^2} = 1$

$\Rightarrow g'(0) = 1$

iii) As $x \to +\infty$

$g'(x) = \frac{1}{(1+x)^2} \to 0$

As $x \to -\infty$

$g'(x) = \frac{1}{(1-x)^2} \to 0$

The gradient indicates that it is vanishing for large values of $|x|$

ADRIAN KUKLA HW2 904056948

Q: 3.1 Gradient Descent

Take derivative of objective function. So,

Let $D(w) = f(w^{(t)}) + \langle w - w^{(t)}, \nabla f(w^{(t)}) \rangle + \frac{\lambda}{2} \| w - w^{(t)} \|^2$

$$\nabla_w D(w) = \nabla f(w^{(t)}) + \lambda(w - w^{(t)}) = 0$$

as $f(w^{(t)})$ is constant, set $w = w^*$

$$\Rightarrow \lambda(w^* - w^{(t)}) = -\nabla f(w^{(t)})$$

$$\Rightarrow w^* = w^{(t)} - \frac{1}{\lambda} \nabla f(w^{(t)})$$

Matches the gradient descent formula for update

$$w^{(t+1)} = w^{(t)} - \eta \nabla f(w^{(t)})$$ where $\eta = \frac{1}{\lambda}$

The update rule corresponds to minimizing the first order Taylor approximation plus an L_2 proximity penalty where the gradient descent step is $\eta = \frac{1}{\lambda}$

λ (regularization term) is the reciprocal of the learning rate in gradient descent. They're inversely proportional

ADRIAN KOKLA HW2 9080 56948

Q 3.2 EC

Prove $\sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle \leq \frac{\|w^*\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|v_t\|^2$

Use telescoping

First,

$\|w^{(t+1)} - w^*\|^2 = \|w^{(t)} - w^* - \eta v_t\|^2$ by definition

$= \|w^{(t)} - w^*\|^2 - 2\eta \langle w^{(t)} - w^*, v_t \rangle + \eta^2 \|v_t\|^2$, by $\|a-b\|^2 = \|a\|^2 - 2\langle a,b \rangle + \|b\|^2$

$\Rightarrow \langle w^{(t)} - w^*, v_t \rangle = \frac{1}{2\eta} \left[\|w^{(t)} - w^*\|^2 - \|w^{(t+1)} - w^*\|^2 \right] + \frac{\eta}{2} \|v_t\|^2$

$\Rightarrow \sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle = \frac{1}{2\eta} \sum_{t=1}^{T} \left(-//- \right) + \frac{\eta}{2} \sum_{t=1}^{T} \|v_t\|^2$

$\Rightarrow \sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle = \frac{1}{2\eta} \left(\|w^{(1)} - w^*\|^2 - \|w^{(T+1)} - w^*\|^2 \right) + \frac{\eta}{2} \sum_{t=1}^{T} \|v_t\|^2$

By assumption $w^{(1)} = 0$ $\Downarrow = \|w^*\|^2$

Also $\|w^{(T+1)} - w^*\|^2 \geq 0$

Hence, $\|w^{(1)} - w^*\|^2 - \|w^{(T+1)} - w^*\|^2 \leq \|w^*\|^2$

$\Rightarrow$ Finally,

$$\sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle \leq \frac{1}{2\eta} \|w^*\|^2 + \frac{\eta}{2} \sum_{t=1}^{T} \|v_t\|^2$$

ADRIAN KOKLA HW2 904056948

Q 3.3 EC

As f is convex we Jensen's inequality where
$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$, $0 \leq \theta \leq 1$

Hence,
$$f(\bar{w}) \leq \frac{1}{T} \sum_{t=1}^{T} f(w^{(t)})$$

$$\Rightarrow f(\bar{w}) - f(w^*) \leq \frac{1}{T} \sum_{t=1}^{T} [f(w^{(t)}) - f(w^*)]$$

$$\Rightarrow f(\bar{w}) - f(w^*) \leq \frac{1}{T} \sum_{t=1}^{T} \langle w^{(t)} - w^*, \nabla f(w^{(t)}) \rangle$$ by first order convexity condition

Part 1 $\nearrow$

Part 2 $\searrow$ Use 3.2 Lemma

$$\|\nabla f(w)\| \leq \rho$$
$$\|\nabla f(w^{(t)})\|^2 \leq \rho^2$$
$$\sum_{t=1}^{T} \|\nabla f(w^{(t)})\|^2 \leq T\rho^2$$

Moreover, $\|w^*\| \leq B \Rightarrow \|w^*\|^2 \leq B^2$

Hence,
$$\frac{1}{T} \sum_{t=1}^{T} \langle w^{(t)} - w^*, \nabla f(w^{(t)}) \rangle \leq \frac{B^2}{2\eta T} + \frac{\eta}{2} \frac{(\cancel{T}\rho^2)}{\cancel{T}}$$

Hence
$$f(\bar{w}) - f(w^*) \leq \frac{B^2}{2\eta T} + \frac{\eta\rho^2}{2}$$

Want to minimize bound by balancing both terms. Hence set
$$\frac{\eta\rho^2}{2} = \frac{B^2}{2\eta T} \Rightarrow \eta^2 = \frac{B^2}{\rho^2 T} \Rightarrow \eta = \frac{B}{\rho\sqrt{T}}$$

$$\Rightarrow \frac{B^2}{2(B/\rho\sqrt{T})T} = \frac{B \cdot \rho}{2\sqrt{T}}, \quad \left(\frac{B}{\rho\sqrt{T}}\right)\frac{\rho^2}{2} = \frac{B\rho}{2\sqrt{T}}$$

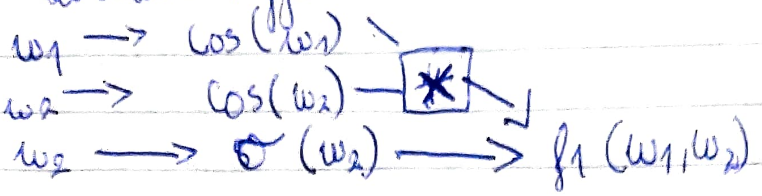
$$\Rightarrow \frac{B\rho}{2\sqrt{T}} + \frac{B\rho}{2\sqrt{T}} = \frac{B\rho}{\sqrt{T}}$$

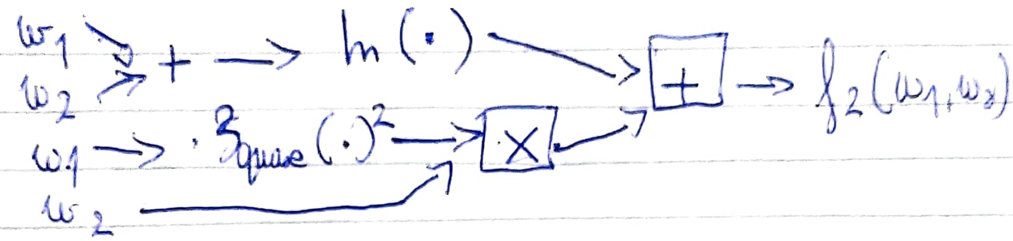
$$\Rightarrow f(\bar{w}) - f(w^*) \leq B\rho/\sqrt{T} = O\left(\frac{1}{\sqrt{T}}\right)$$ Is the convergence rate

ADRIAN KURLA HW2 904056928

Q4 Automatic Differentiation

a)





$$f_1(1,2) \Rightarrow \cos(1)\cos(2) = 0.54 \times -0.416 = -0.225$$

$$\sigma(2) = \frac{1}{1+e^{-2}} \approx 0.8807$$

$$f_1(1,2) = -0.225 + 0.8807 \approx 0.656$$

$$f_2(1,2) = \ln 3 + 2 \approx 3.0986$$

$$f(1,2) \approx (0.656, 3.0986)$$

b) Using Excel find

$$J(w) = \begin{pmatrix} \frac{\partial f_1}{\partial w_1} & \frac{\partial f_1}{\partial w_2} \\ \frac{\partial f_2}{\partial w_1} & \frac{\partial f_2}{\partial w_2} \end{pmatrix} \text{ where}$$

$$\frac{\partial f_i}{\partial w_j} \approx \frac{f_i(w_1 + \Delta_j, w_2 + \Delta_j) - f_i(1,2)}{0.01}$$

$$\frac{\partial f_1}{\partial w_1} \approx \frac{f_1(1.01, 2) - f_1(1,2)}{0.01} = 0.35$$

$$\frac{\partial f_1}{\partial w_2} \approx \frac{f_1(1, 2.01) - f_1(1,2)}{0.01} \approx -0.39$$

$$\frac{\partial f_2}{\partial w_1} \approx \frac{f_2(1.01, 2) - f_2(1,2)}{0.01} \approx 4.35$$

$$\frac{\partial f_2}{\partial w_2} \approx \frac{f_2(1, 2.01) - f_2(1,2)}{0.01} \approx 1.33$$

$$J(w) = \begin{bmatrix} 0.35 & -0.39 \\ 4.35 & 1.33 \end{bmatrix}$$

ADRIAN KUKLA HW2 904056948

Q4
c)

$$\frac{\partial f_1}{\partial w_1} = -\sin(w_1)\cos(w_2) = 0.35$$

$$\frac{\partial f_1}{\partial w_2} = -\cos(w_1)\sin(w_2) \cancel{+} \sigma(w_2)(1-\sigma(w_2)) = -0.39$$

$$\frac{\partial f_2}{\partial w_1} = \frac{1}{w_1+w_2} + 2w_1w_2 = 4.33$$

$$\frac{\partial f_2}{\partial w_2} = \frac{1}{w_1+w_2} + w_1^2 = 1.33$$

$$J(w) = \begin{bmatrix} 0.35 & -0.39 \\ 4.33 & 1.33 \end{bmatrix}$$

d) Backward mode yields the same result as the forward mode for a function $f: \mathbb{R}^2 \to \mathbb{R}^2$. It uses the chain rule to differentiate w.r.t. intermediate functions to derive same result

$$J(w) = \begin{bmatrix} 0.35 & -0.39 \\ 4.33 & 1.33 \end{bmatrix}$$

e) Yes I love it - automatic differentiation eliminates manual derivative calculations.

ADRIAN KUKLA HW2 9040569228

Q5 Convolutions

a) Need to show $SC = CS$. Show by inspection

$$SC = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \\ 1 & \ddots & & & 0 \\ & \ddots & & & 0 \\ & & \ddots & & \vdots \\ & & & 1 & 0 \end{pmatrix} \begin{pmatrix} a_0 & a_{n-1} & a_{n-2} & \cdots & a_1 \\ a_1 & a_0 & a_{n-1} & & a_2 \\ a_2 & \ddots & & \ddots & \vdots \\ \vdots & & \ddots & & \vdots \\ a_{n-1} & a_{n-2} & \cdots & & a_0 \end{pmatrix} = \begin{pmatrix} a_{n-1} \\ a_{n-1} \\ \vdots \\ a_{n-1} \end{pmatrix}$$

Diagonals multiply a set of 1's and a_{n-1}

$$CS = \begin{pmatrix} a_0 & a_{n-1} & a_{n-2} & \cdots & a_1 \\ a_1 & a_0 & a_{n-1} & & a_2 \\ a_2 & \ddots & & \ddots & \vdots \\ \vdots & & \ddots & & \vdots \\ a_{n-1} & a_{n-2} & \cdots & & a_0 \end{pmatrix} \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & \cdots & & \\ & \ddots & & & \\ & & \ddots & & \\ & & & 1 & 0 \end{pmatrix} = \begin{pmatrix} a_{n-1} \\ a_{n-1} \\ \vdots \\ a_{n-1} \end{pmatrix}$$

Moreover $C_a(Sx) = S(C_a)x \;\forall x \Rightarrow C_a S = S C_a$

Shifting input and then convolving is equivalent to convolving & then shifting

b) Forward direction $\Rightarrow$ if an operation is a circular convolution then it is represented by a circulant matrix C_a which commutes with S i.e. $SC = CS$

Reverse direction $\Rightarrow$ Let's say L is shift equivariant so

$$L(Sx) = S(Lx)$$

Because L is linear it implies it is a matrix M (linearity equation)

Since L is shift equivariant then $MS = SM$
$\quad\;\; {}^{\shortparallel}M$

However, $M = a_0 I + \alpha_1 S + \alpha_2 S^2 + \ldots + \alpha_{n-1} S^{n-1}$

Since a matrix that commutes with S is a set of polynomials of S and polynomials of S are circulant matrices,

Hence M is a circulant matrix s.th.

$$M(Sx) = S(Mx) \quad \forall x$$

ADRIAN KOKLA HW2 904056948

Q5

c) Means that CNNs are very good at analysing computer vision related tasks. So convolutions are shift equivariant and hence this means that convolutional layers can detect features independent of their position in spatial or time space. This avoids redundant parameter learning across positions.

Equivariance ensures robustness to translations which improves performance on tasks where the alignment of inputs varies like motion detection in videos or object recognition.

Also CNNs can reduce number of parameters compared to just a fully connected layer thanks to the shift equivariance property.

For 2D images 2D convolutions are used

For 3D videos 3D convolutions are used.

ADRIAN KURLA HW2 9040568948

Q6

$A_1(w) = \frac{1}{2}(w-2)^2 \Rightarrow \nabla A_1(w) = w-2$

$A_2(w) = \frac{1}{2}(w+1)^2 \Rightarrow \nabla A_2(w) = w+1$

$w_{new} = w_{old} - \eta \nabla [\text{sampled term}]$

$A_1(w)$ w_{new} moves in direction $-(w-2)$

$A_2(w)$ w_{new} moves in direction $-(w+1)$

Select $w=0$

A_1 at $w=0$ $\nabla A_1(0) = -2$

$\Rightarrow w_{new} = 0 - \eta(-2) = 2\eta$

Evaluate $f(2\eta) < f(0)$

$f(w) = \frac{1}{2}[w^2 - 4w + 4 + w^2 + 2w + 1] = w^2 - w + 2.5$

$f(2\eta) = 4\eta^2 - 2\eta + 2.5$ $f(0) = 2.5$

if $\eta < 1/2$

$f(2\eta) - f(0) = 4\eta^2 - 2\eta < 0$

Hence $f(2\eta) < f(0)$ hence f decreases

A_2 at $w=0$ $\nabla A_2(0) = 1$

$\Rightarrow w_{new} = 0 - \eta(1) = -\eta$

$f(-\eta) = \eta^2 + \eta + 2.5$

$f(-\eta) - f(0) = \eta^2 + \eta + 2.5 - 2.5 = \eta^2 + \eta > 0$

so f increases

Hence for $w=0$ if A_2 is sampled, the next iteration will increase the overall function f.

Hence.

SGD is NOT guaranteed to decrease overall loss function.

Adrian Kukla

CS7643 HW2

Question 7

a)

Key contributions

Paper shows that deep neural networks can perfectly fit random labels.

Shows that explicit regularization like dropout or weight decay are neither necessary nor sufficient for explaining generalization. The paper suggests that is it **not** likely that regularization techniques are the main reason for generalization.

Shows that any depth 2-layer networks of linear size with ReLU activations can represent any labelling of the training data (any function) which shows that deep neural nets have very high capacity for learning.

Shows that SGD implicitly regularizes solutions which leads to generalization. SGD even with unchanged parameters can optimize weights to fit random patterns perfectly even when there is no relationship between labels and the associated images.

Strengths

Authors of the paper refer to works of multiple other authors/papers to provide some context.

Work of authors challenges conventional wisdom rather than conforming to the status quo – which can spur more research and innovations.

Study does conduct robust experiments on a number of different types of neural nets with and without regularization to evaluate their hypothesis. Combines empirical findings with ground neural network theory.

Weaknesses

No explanation about why some networks generalize better than others.

Doesn't offer any solutions or techniques to improve understanding of generalization.

Doesn't offer alternatives to classical and established measures.

b)

Personal takeaways

My understanding was that overfitting tends to occur when number of parameters is higher than number of samples – however this paper provides a counter to my previous understanding. Moreover, I thought that explicit regularization techniques like dropout or weight decay is necessary for generalization, whereas the paper shows that they're not. Changing the model architecture appropriately can vastly improve generalization without use of explicit regularization terms.

Interesting that the paper ruled out common measures like VC dimensions, Rademacher complexity as potential explanations for generalization performance. The paper suggests that new theoretical frameworks should be explored which are outside of the commonly accepted measures.

Moreover, potential future research would be to evaluate what common properties of models that were trained by SGD to analyse how these models generalize well without the need for explicit regularization.