CSC413: Homework 1

Tianyu Du (1003801647)

January 24, 2020

1 Hard-Coding Networks

1.1 Verify Sort

Soln. The first layer performs pairwise comparison to construct indicators $\mathbb{1}\{x_1 \leq x_2\}$, $\mathbb{1}\{x_2 \leq x_3\}$, and $\mathbb{1}\{x_3 \leq x_4\}$. The second layer performs an `all()` operation on indicators from the previous layer.

$$\mathbf{W}^{(1)} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix} \tag{1.1}$$

$$\mathbf{b}^{(1)} = \begin{pmatrix} 0 & 0 & 0 \end{pmatrix} \tag{1.2}$$

So that

$$\varphi(\mathbf{h}) = \varphi(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}) = \varphi \begin{pmatrix} x_2 - x_1 \\ x_3 - x_2 \\ x_4 - x_3 \end{pmatrix} = \begin{pmatrix} \mathbb{1}\{x_2 \geq x_1\} \\ \mathbb{1}\{x_3 \geq x_2\} \\ \mathbb{1}\{x_4 \geq x_3\} \end{pmatrix} \tag{1.3}$$

$$\mathbf{w}^{(2)} = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \tag{1.4}$$

$$b^{(2)} = -0.5 \tag{1.5}$$

Such that $y = 1$ if and only if all components of $\mathbf{h}$ are ones, i.e., the list is sorted. ■

1.2 Perform Sort

Soln. ■

1.3 Universal Approximation Theorem

1.3.1

Soln. To avoid over-using of notations, let $\varphi(y) := \mathbb{1}\{y > 0\}$ denote the activation function.

$$n = 2 \tag{1.6}$$
$$\mathbf{W}_0 = (1, -1) \tag{1.7}$$
$$\mathbf{b}_0 = (-a, b) \tag{1.8}$$
$$\mathbf{W}_1 = (1, 1) \tag{1.9}$$
$$\mathbf{b}_1 = -0.5 \tag{1.10}$$

Justification:

$$\varphi(\mathbf{h}) = \varphi((x - a, b - x)) \tag{1.11}$$
$$= (\mathbb{1}\{x - a > 0\}, \mathbb{1}\{b - x > 0\}) \tag{1.12}$$
$$= (\mathbb{1}\{x > a\}, \mathbb{1}\{x < b\}) \tag{1.13}$$
$$\varphi(\mathbf{W}_1\varphi(\mathbf{h}) + \mathbf{b}_1) = \mathbb{1}\{\mathbb{1}\{x > a\} + \mathbb{1}\{x < b\} - 0.5\} \tag{1.14}$$
$$= \mathbb{1}\{x > a\} \wedge \mathbb{1}\{x < b\} \tag{1.15}$$
$$= \mathbb{1}\{a < x < b\} \tag{1.16}$$

■

1.3.2

Soln. Let $\delta \in (0, 1)$ denote the ratio parameter, a higher value of δ results in a finer approximation is. In this example, take $\delta = \frac{9}{10}$.
Without loss of generality, assume the region I on which function f is defined on to be symmetric across zero.
Let $I = [-1, 1]$, given f is symmetric, $f(-\delta) = f(\delta)$.
Define:

$$\hat{f}_1(x) = \hat{f}_0(x) + g(f(\delta), -\delta, \delta, x) \tag{1.17}$$

Note that

$$||f - \hat{f}_1|| = \int_{-1}^{1} \left| f(x) - \hat{f}_1(x) \right| \, dx \tag{1.18}$$
$$= \int_{-1}^{-\delta} |f(x)| \, dx + \int_{-\delta}^{\delta} \left| f(x) - \hat{f}_1(x) \right| \, dx + \int_{\delta}^{1} |f(x)| \, dx \tag{1.19}$$

Given that $\forall x \in (-\delta, \delta),\ f(x) > f(-\delta) = f(\delta) > 0$, it follows

$$\int_{-\delta}^{\delta} \left| f(x) - \hat{f}_1(x) \right| \, dx = \int_{-\delta}^{\delta} f(x) - \hat{f}_1(x) \, dx \tag{1.20}$$
$$= \int_{-\delta}^{\delta} f(x) \, dx - \int_{-\delta}^{\delta} \hat{f}_1(x) \, dx \tag{1.21}$$

Also, $\int_{-\delta}^{\delta} \hat{f}_1(x)\ dx > 0$ provided $\delta \neq 0$. Therefore,

$$\int_{-\delta}^{\delta} \left| f(x) - \hat{f}_1(x) \right|\ dx < \int_{-\delta}^{\delta} f(x)\ dx \tag{1.22}$$
$$= \int_{-\delta}^{\delta} |f(x)|\ dx \tag{1.23}$$

Therefore,

$$||f - \hat{f}_1|| = \int_{-1}^{-\delta} |f(x)|\ dx + \int_{-\delta}^{\delta} \left| f(x) - \hat{f}_1(x) \right|\ dx + \int_{\delta}^{1} |f(x)|\ dx \tag{1.24}$$
$$< \int_{-1}^{-\delta} |f(x)|\ dx + \int_{-\delta}^{\delta} |f(x)|\ dx + \int_{\delta}^{1} |f(x)|\ dx \tag{1.25}$$
$$= \int_{-1}^{1} |f(x) - 0|\ dx \tag{1.26}$$
$$= \int_{-1}^{1} \left| f(x) - \hat{f}_0(x) \right|\ dx \tag{1.27}$$
$$= ||f(x) - \hat{f}_0(x)|| \tag{1.28}$$

Therefore,

$$||f(x) - \hat{f}_1(x)|| < ||f(x) - \hat{f}_0(x)|| \tag{1.29}$$

■

1.3.3

Soln. **Algorithm:**

(i) Divide $I = [-1, 1]$ into $N + 2$ sub-intervals with equal length, such that

$$I_i := \left[-1 + \frac{i}{N+2}, -1 + \frac{i+1}{N+2} \right] \quad \forall\ i \in \{1, 2, \cdots, N\} \tag{1.30}$$

Note that the first and last sub-intervals are not used to construct g_i.

(ii) For each i, define

$$h_i := \min_{x \in I_i} f(x) \tag{1.31}$$
$$a_i := -1 + \frac{i}{N+2} \tag{1.32}$$
$$b_i := -1 + \frac{i+1}{N+2} \tag{1.33}$$

Because $f(x) \geq 0 \ \forall x \in I$.
By the definition of $g_i(x)$, it can be shown that[1]

$$f(x) \geq f_i(x) \ \forall i \in \{1, 2, \cdots, N\} \ \forall x \in \bigcup_{i=1}^{N} (a_i, b_i) \tag{1.34}$$

Further,

$$f(x) = f_i(x) \quad \forall i \in \{1, 2, \cdots, N\} \ \forall x \in \left[-1, -1 + \frac{1}{N+2}\right) \bigcup \left(1 - \frac{1}{N+2}, 1\right] \tag{1.35}$$

Define

$$\mathcal{K} := \left[-1, -1 + \frac{1}{N+2}\right) \bigcup \left(1 - \frac{1}{N+2}, 1\right] \bigcup \left(\bigcup_{i=1}^{N} (a_i, b_i)\right) \tag{1.36}$$

Note that the set $I \backslash \mathcal{K}$ consists of all boundary points between consecutive sub-intervals. There are only finitely many such points, therefore $I \backslash \mathcal{K}$ has measure zero, and

$$\int_I \left| f(x) - \hat{f}_i(x) \right| \ dx = \int_{\mathcal{K}} \left| f(x) - \hat{f}_i(x) \right| \ dx \tag{1.37}$$

And I've shown that for every i and every $x \in \mathcal{K}$, $f(x) \geq f_i(x)$. Consequently,

$$\begin{aligned}
\int_I \left| f(x) - \hat{f}_i(x) \right| \ dx &= \int_{\mathcal{K}} \left| f(x) - \hat{f}_i(x) \right| \ dx \text{ (removing measure zero set.)} && (1.38)\\
&= \int_{\mathcal{K}} f(x) - \hat{f}_i(x) \ dx && (1.39)\\
&= \int_I f(x) - \hat{f}_i(x) \ dx \text{ (adding back the measure zero set.)} \quad (\dagger) && (1.40)
\end{aligned}$$

Define $\hat{f}_0(x) = 0$ and let $i \in \{1, 2, \cdots, N\}$,

$$\begin{aligned}
||f - \hat{f}_{i+1}|| &= \int_{-1}^{1} \left| f(x) - \hat{f}_{i+1}(x) \right| \ dx && (1.41)\\
&= \int_{-1}^{1} f(x) - \hat{f}_{i+1}(x) \ dx \text{ by } (\dagger) && (1.42)\\
&= \int_{-1}^{-1\frac{i+1}{N}} f(x) - \hat{f}_{i+1}(x) \ dx + \int_{-1+\frac{i+1}{N}}^{-1+\frac{i+2}{N}} f(x) - \hat{f}_{i+1}(x) \ dx + \int_{-1+\frac{i+2}{N}}^{1} f(x) - \hat{f}_{i+1}(x) \ dx && (1.43)
\end{aligned}$$

[1] I am excluding those boundary points between consecutive sub-intervals, because at those points, the value of f_i spikes due to duplicate counts of indicator functions. However, while doing integral, this does not matter as the set of boundary points has measure zero.

Further, by construction, $\hat{f}_{i+1}(x) = \hat{f}_i(x)\ \forall x \notin [a_{i+1}, b_{i+1}]$. Therefore,

$$\begin{aligned}
||f - \hat{f}_{i+1}|| &= \int_{-1}^{a_{i+1}} f(x) - \hat{f}_i(x)\ dx + \int_{a_{i+1}}^{b_{i+1}} f(x) - \hat{f}_{i+1}(x)\ dx + \int_{b_{i+1}}^{1} f(x) - \hat{f}_i(x)\ dx && (1.44) \\
&= \int_{-1}^{a_{i+1}} f(x) - \hat{f}_i(x)\ dx + \int_{a_{i+1}}^{b_{i+1}} f(x) - \hat{f}_i(x) - g(h_{i+1}, a_{i+1}, b_{i+1}, x)\ dx + \int_{b_i}^{1} f(x) - \hat{f}_i(x)\ dx && (1.45) \\
&= \int_{-1}^{a_{i+1}} f(x) - \hat{f}_i(x)\ dx + \int_{a_{i+1}}^{b_{i+1}} f(x) - \hat{f}_i(x)\ dx + \int_{b_{i+1}}^{1} f(x) - \hat{f}_i(x)\ dx - \int_{a_{i+1}}^{b_{i+1}} g(h_{i+1}, a_{i+1}, b_{i+1}, x)\ dx && (1.46) \\
&= \int_{-1}^{1} f(x) - \hat{f}_i(x)\ dx - \int_{a_{i+1}}^{b_{i+1}} g(h_{i+1}, a_{i+1}, b_{i+1}, x)\ dx && (1.47) \\
&= ||f - \hat{f}_i|| - \int_{a_{i+1}}^{b_{i+1}} g(h_{i+1}, a_{i+1}, b_{i+1}, x)\ dx && (1.48)
\end{aligned}$$

Note that for every i, for every $x \in [a_i, b_i]$, $g(h_i, a_i, b_i, x) > 0$. Therefore, $\int_{a_i}^{b_i} g(h_i, a_i, b_i, x)\ dx > 0$. Hence,

$$\begin{aligned}
||f - \hat{f}_{i+1}|| &= ||f - \hat{f}_i|| - \int_{a_{i+1}}^{b_{i+1}} g(h_{i+1}, a_{i+1}, b_{i+1}, x)\ dx && (1.49) \\
&> ||f - \hat{f}_i|| && (1.50)
\end{aligned}$$

■

1.3.4

Soln. Not required. ■

2 Backprop

2.1 Computational Graph

2.1.1

Soln. TODO: *Add graph* ■

2.1.2

Soln.

$$\begin{aligned}
\overline{\mathbf{x}} &= \overline{\mathbf{z}}\frac{\partial \mathbf{z}}{\partial \mathbf{x}} && (2.1)\\
&= \overline{\mathbf{z}}\mathbf{W}^{(1)} && (2.2)\\
&= \overline{\mathbf{h}}\frac{\partial \mathbf{h}}{\partial \mathbf{z}}\mathbf{W}^{(1)} && (2.3)\\
&= \overline{\mathbf{h}}\mathbb{1}\{\mathbf{z} \geq 0\}\mathbf{W}^{(1)} && (2.4)\\
&= \left(\overline{\mathcal{R}}\frac{\partial \mathcal{R}}{\partial \mathbf{h}} + \overline{\mathbf{y}}\frac{\partial \mathbf{y}}{\partial \mathbf{h}}\right)\mathbb{1}\{\mathbf{z} \geq 0\}\mathbf{W}^{(1)} && (2.5)\\
&= \left(\overline{\mathcal{R}}\mathbf{r}^T + \overline{\mathbf{y}}\mathbf{W}^{(2)}\right)\mathbb{1}\{\mathbf{z} \geq 0\}\mathbf{W}^{(1)} && (2.6)\\
&= \left(\mathbf{r}^T + \overline{\mathbf{y}'}\frac{\partial \mathbf{y}'}{\partial \mathbf{y}}\mathbf{W}^{(2)}\right)\mathbb{1}\{\mathbf{z} \geq 0\}\mathbf{W}^{(1)} && (2.7)\\
&= \left(\mathbf{r}^T + \overline{\mathbf{y}'}\texttt{softmax}'(\mathbf{y})\mathbf{W}^{(2)}\right)\mathbb{1}\{\mathbf{z} \geq 0\}\mathbf{W}^{(1)} && (2.8)\\
&= \left(\mathbf{r}^T + \overline{\mathcal{S}}\frac{\partial \mathcal{S}}{\partial \mathbf{y}'}\texttt{softmax}'(\mathbf{y})\mathbf{W}^{(2)}\right)\mathbb{1}\{\mathbf{z} \geq 0\}\mathbf{W}^{(1)} && (2.9)\\
&= \left(\mathbf{r}^T + \mathbf{e}_k\ \texttt{softmax}'(\mathbf{y})\mathbf{W}^{(2)}\right)\mathbb{1}\{\mathbf{z} \geq 0\}\mathbf{W}^{(1)} && (2.10)
\end{aligned}$$

where $\mathbf{e}_k$ denotes the one-hot vector in $\mathbb{R}^M$ in which the k^{th} element is one. ■

2.2 Vector-Jacobean Product (VJPs)

2.2.1

2.2.2

2.2.3

3 Linear Regression

3.1 Driving the Gradient

Soln.

$$\begin{aligned}
\frac{d}{d\hat{\mathbf{w}}}\frac{1}{n}(X\hat{\mathbf{w}} - \mathbf{t})^2 &= \frac{d}{d\hat{\mathbf{w}}}\frac{1}{n}||X\hat{\mathbf{w}} - \mathbf{t}||_2^2 && (3.1)\\
&= \frac{2}{n}(X\hat{\mathbf{w}} - \mathbf{t})^T X && (3.2)\\
\implies \nabla_{\mathbf{w}}\mathcal{J} &= \frac{2}{n}X^T(X\hat{\mathbf{w}} - \mathbf{t}) && (3.3)
\end{aligned}$$

■

3.2 Under-parameterized Model

3.2.1

Soln. Assume $d < n$ so that $X^T X$ is invertible. The gradient descent algorithm converges when the gradient equals zero:

$$\frac{2}{n}(X\hat{\mathbf{w}} - \mathbf{t})^T X = 0 \tag{3.4}$$
$$\implies (X\hat{\mathbf{w}} - \mathbf{t})^T X = 0 \tag{3.5}$$
$$\implies X^T(X\hat{\mathbf{w}} - \mathbf{t}) = 0^T \tag{3.6}$$
$$\implies X^T X\hat{\mathbf{w}} - X^T\mathbf{t} = 0^T \tag{3.7}$$
$$\implies X^T X\hat{\mathbf{w}} = X^T\mathbf{t} \tag{3.8}$$
$$\implies \hat{\mathbf{w}} = (X^T X)^{-1} X^T\mathbf{t} \tag{3.9}$$

■

3.2.2

Soln. Let $\mathbf{x} \in \mathbb{R}^d$, note that $(X^T X)^{-1}$ is symmetric. Assuming target $\mathbf{t}$ is generated by a linear process, then $\mathbf{t} = X\mathbf{w}^*$. Immediately, $\mathbf{t}^T = \mathbf{w}^{*T} X^T$.

$$(\mathbf{w}^{*T}\mathbf{x} - \hat{\mathbf{w}}^T\mathbf{x})^2 = (\mathbf{w}^{*T}\mathbf{x} - [(X^T X)^{-1}X^T\mathbf{t}]^T\mathbf{x})^2 \tag{3.10}$$
$$= (\mathbf{w}^{*T}\mathbf{x} - \mathbf{t}^T X(X^T X)^{-1}\mathbf{x})^2 \tag{3.11}$$
$$= (\mathbf{w}^{*T}\mathbf{x} - \mathbf{w}^{*T} X^T X(X^T X)^{-1}\mathbf{x})^2 \tag{3.12}$$
$$= (\mathbf{w}^{*T}\mathbf{x} - \mathbf{w}^{*T}\mathbf{x})^2 \tag{3.13}$$
$$= 0 \tag{3.14}$$

■

3.3 Over-parameterized Model: 2D Example

3.3.1

Soln. To minimize the empirical risk minimizer,

$$\min_{w_1, w_2} (w_1 x_1 + w_2 x_2 - t_1)^2 \tag{3.15}$$
$$\text{equivalently,} \ \min_{w_1, w_2} (2w_1 + w_2 - 2)^2 \tag{3.16}$$

Any pair of (w_1, w_2) satisfying

$$2w_1 + w_2 - 2 = 0 \quad (\dagger) \tag{3.17}$$

attains the minimum level of empirical risk (zero). Equivalently, any $\hat{\mathbf{w}}$ on the line

$$\hat{\mathbf{w}} = \begin{pmatrix} 0 \\ 2 \end{pmatrix} + t \begin{pmatrix} 1 \\ -2 \end{pmatrix} \quad \text{for } t \in \mathbb{R} \tag{3.18}$$

satisfies (†). Therefore, there are infinitely many empirical risk minimizers.
Equivalently, the collection of solution is

$$w_2 = -2w_1 + 2 \tag{3.19}$$

■

3.3.2

Soln. From the fist part of this question we know that

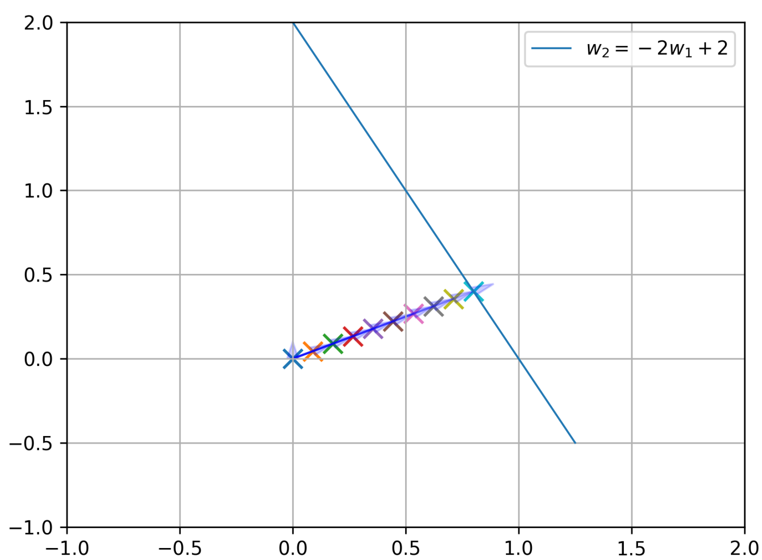
$$\begin{aligned}
\nabla_{\mathbf{w}}\mathcal{J} &= \frac{2}{n}(X\hat{\mathbf{w}} - \mathbf{t})^T X && (3.20)\\
&= \frac{2}{1}\left[\begin{pmatrix}2\\1\end{pmatrix}\cdot\begin{pmatrix}0\\0\end{pmatrix} - 2\right]\begin{pmatrix}2\\1\end{pmatrix} && (3.21)\\
&= \begin{pmatrix}-4\\-2\end{pmatrix} && (3.22)
\end{aligned}$$

The unit-norm gradient is

$$\widehat{\nabla_{\mathbf{w}}\mathcal{J}} = \begin{pmatrix}-\frac{2\sqrt{5}}{5}\\ -\frac{\sqrt{5}}{5}\end{pmatrix} \tag{3.23}$$

The direction (gradient) does not change along the trajectory. Ultimately, the gradient descend algorithm converges to

$$\hat{\mathbf{w}}^* = \left(\frac{4}{5}, \frac{2}{5}\right) \tag{3.24}$$



■

3.3.3

Soln. Let $\hat{\mathbf{w}}^*$ denote the solution found using gradient descent. Note that the line of solution can be written parametrically as

$$\hat{\mathbf{w}} = \begin{pmatrix} 0 \\ 2 \end{pmatrix} + t \begin{pmatrix} 1 \\ -2 \end{pmatrix} \tag{3.25}$$

Notice that the path of $\hat{\mathbf{w}}_t$ during the process of gradient descend, the path is perpendicular to the direction of the line of solutions. Therefore, the attained $\hat{\mathbf{w}}^*$ is the one nearest to the initial point, $\hat{\mathbf{w}}_0$. Here $\hat{\mathbf{w}}_0 = \mathbf{0}$, and the solution is therefore the one has the smallest Euclidean norm. TODO: *Formalize this proof.* ■

3.4 Overparameterized Model: General Case

3.4.1

Proof. ■

3.4.2

Proof. ■

3.5 Benefit of Overparameterization

3.5.1

Soln. ■

3.5.2

Soln. Not required. ■