CSE 599 D1 (Neural Networks Theory): Homework #2

Due on March, 3rd, 2021 at 23:59

Prof. Simon Du

Alexey Sholokhov

Problem 1

Solution for 1.1 From the lecture on February 1st we know that the gradient flow yields the following dynamics when applied to a least-squares fitting problem:

$$\frac{\mathrm{d}}{\mathrm{d}t}(\frac{1}{2}\|u(t)-y\|_2^2) = -\frac{1}{n}(u(t)-y)^T H(t)(u(t)-y) \tag{1}$$

where $u(t)$ is a model with parameters w_t, y is the vector of target variable, and

$$[H(t)]_{ij} \triangleq \left\langle \frac{\partial [u(t)]_i}{\partial w}, \frac{\partial [u(t)]_j}{\partial w} \right\rangle \tag{2}$$

In this problem, our model $u(t)$ is $X^T w_t$, hence $H = X^T X \in \mathbb{R}^{n\times n}$ is a full-rank matrix that does not depend on t and has a positive minimum eigenvalue. It exactly meets the assumptions of the "Consequence in Training" corollary from the lecture notes where we show that in such case the loss goes to zero exponentially fast. The convergence speed is determined by the lower-bound λ_0 on the smallest eigenvalue of the matrix H:

$$\frac{1}{2}\|u(t)-y\|_2^2 \leq e^{-\frac{\lambda_0 t}{2}} \tag{3}$$

□

Solution for 1.2 Our loss function is defined as

$$L(w_t) = \frac{1}{2n}\|X^T w_t - y\|_2^2 \tag{4}$$

The gradient of $L(w)$ is

$$\nabla_w L(w_t) = \frac{1}{2} X(X^T w_t - y) \tag{5}$$

Hence, the gradient flow equation for this problem is

$$\frac{\mathrm{d}}{\mathrm{d}t}(w_t) = -\frac{1}{n} X^T X w_t + \frac{1}{n} X y \tag{6}$$

Due to the principle of superposition, a solution for this problem is a sum of solutions for its homogeneous and inhomogeneous parts:

$$w(t) = w_H(t) + w_{NH}(t) \tag{7}$$

From taking ODE classes we know that the general solution for the homogeneous equation $\dot{w} = Aw$ lies in the basis of the singular vectors of the matrix A. For our problem, $A = \frac{1}{n} X X^T$.
There are couple things to notice about the SVD decomposition for XX^T. Let's say that $X = U\Sigma V$. First, we know that taking a power of the matrix does not change the content of its singular basis (columns of U):

$$X = U\Sigma V^* \implies XX^T = U\Sigma^2 U^* \tag{8}$$

Second, we know that for an SVD of X we can write that

$$XV = U\Sigma \tag{9}$$

Since Σ is a diagonal matrix non-zeros in first n diagonal entries, we can surely say that the first n columns of U have to belong to the column-span of X. In general, it's not true about the rest $d-n$ columns of U: they are chosen so that U is an orthonormal matrix, and the equation above holds for these vectors not because

Problem 1 continued on next page...

they belong to the column-span of X, but because the respective diagonal entries in Σ are zero, and so this part of U cancels out. Keeping this in mind, we split the solution $w_H(t)$ for the homogeneous part into two sums:

$$w_H(t) = \frac{1}{n}\sum_{i=1}^{n} C_i u_i e^{-\lambda_i^2 t} + \frac{1}{n}\sum_{j=n+1}^{d} C_j u_j \cancelto{1}{e^{-\lambda_j^2 t}} \tag{10}$$

where C_i and C_j are constant coefficients. The non-homogeneous part of this ODE is separable by its coordinates, so we obtain the general solution by integrating coordinate-wise:

$$w_{NH}(t) = \frac{t}{n} Xy + D \tag{11}$$

where D is an unknown constant vector. The overall general solution is

$$w(t) = w_H(t) = \frac{1}{n}\sum_{i=1}^{n} C_i u_i e^{-\lambda_i^2 t} + \frac{t}{n} Xy + D' \tag{12}$$

where $D' = D + \frac{1}{n}\sum_{j=n+1}^{d} C_j u_j$ is an undefined fixed vector. The first two summands in the solution belong to the column-span of X, as we discussed above. We show that D' also belongs to this span using the initial condition:

$$0 = w(0) = \frac{1}{n}\sum_{i=1}^{n} C_i u_i + D' \tag{13}$$

Since $\{u_i\}_{i=1}^n$ belong to the column-span of X, D' also must belong to this span:

$$D' = -\frac{1}{n}\sum_{i=1}^{n} C_i u_i \tag{14}$$

Hence, the whole solution $w(t)$ lies in $\{x_1, \ldots, x_n\}$. □

Solution for 1.3 From 1.2, we know that w_t always lies in the span of $\{x_1, \ldots, x_n\}$. It means that there are coefficients $\{c_1(t), \ldots, c_n(t)\}$ such that

$$w(t) = \sum_{i=1}^{n} c_i(t) x_i \tag{15}$$

This is also true for the limiting case $t \to \infty$, that implies

$$w^\infty = \sum_{i=1}^{n} c_i^\infty x_i = X c^\infty \tag{16}$$

From 1.1 we know that $\lim_{t\to\infty} L(w(t)) = 0$, which means that the limiting solution fits the data precisely:

$$x_i^T w^\infty = y_i, \text{ for all } i \in 1, \ldots, n \tag{17}$$

Combining these two statements we're getting that

$$X^T X c^\infty = y \tag{18}$$

Since $X^T X$ is a full-rank matrix, we have that its inverse exists and

$$c^\infty = (X^T X)^{-1} y \tag{19}$$

Problem 1 continued on next page...

and so the limiting solution is

$$w^{\infty} = X(X^T X)^{-1} y \tag{20}$$

Now let's save this result and then consider the second optimization problem independently of the results above:

$$\begin{aligned} &\min_{w} \|w\| \\ &\text{s.t. } X^T w = y \end{aligned} \tag{21}$$

Since $\|w\| = \|w - 0\|$ we immediately notice that this is a projection problem, were we're looking for a projection of the origin onto the hyperplane $X^T w = y$. From a linear algebra class, we know that this projection obeys the normal equation:

$$X(X^T w - y) = 0 \tag{22}$$

The solution is unique in the minimal-norm sense and can be found via pseudo-inverse:

$$w^* = (XX^T)^{\dagger} X y \tag{23}$$

In the class (Feb, 1st), we showed that $(X^T X)^{-1} X^T = X^T (XX^T)^{\dagger}$. Taking the transpose of both sides and noticing that $X^T X$ and XX^T are symmetric matrices and so their inverses and pseudo-inverses are too, we get $X(X^T X)^{-1} = (XX^T)^{\dagger} X$. This implies that $w^* = w^{\infty}$ which concludes the proof. □

Problem 2

Solution for 2.1 Let's fix some activation pattern a' and consider $x, y \in S_{a'}$, and $\alpha, \beta \in \mathbb{R}$ such that $\alpha x + \beta y \in S_{a'}$. We will prove by induction that $x_h(x) = A_h(x) W_h x_{h-1}(x)$ is a linear function when restricted to $S_{a'}$.
Let's consider the base:

$$\begin{aligned} x_1(\alpha x + \beta y) &= A(\alpha x + \beta y) W_1 (\alpha x + \beta y) = \alpha A(\alpha x + \beta y) W_1 x + \beta A(x + y) W_1 y \\ &= \alpha A(x) W_1 x + \beta A_1(y) W_1 y = \alpha x_1(x) + \beta x_1(y) \end{aligned} \tag{24}$$

Now, assuming that x_{h-1} is a linear function we show that x_h is also a linear function:

$$\begin{aligned} x_h(\alpha x + \beta y) &= A_h(\alpha x + \beta y) W_h x_{h-1}(\alpha x + \beta y) = \\ &= A_h(\alpha x + \beta y) W_h (\alpha x_{h-1}(x) + \beta x_{h-1}(y)) \\ &= \alpha A_h(x) W_h x_{h-1}(x) + \beta A_h(y) W_h x_{h-1}(y) \\ &= \alpha x_h(x) + \beta x_h(y) \end{aligned} \tag{25}$$

where we used the fact that $A_h(\alpha x + \beta y) = A_h(x) = A_h(y)$ which is true when the inputs are restricted to $S_{a'}$. Finally, since we proved that $x_h(x)$ is linear, the linearity of $f(x) = W_{H+1} x_H(x)$ follows from the linearity of the scalar product. □

Solution for 2.2 There are two approaches to prove this statement: a simple one with a looser bound and a more involved one with a better bound. We will give the first one and reference the second one.
Approach 1: as we've proven in 2.1, $f(x)$ is a linear function on every region $S_{a'}$. Every $x \in \mathbb{R}^d$ belongs to some $S_{a'}$ according to the activation pattern that it produces for a given neural network. It means that all $\mathbb{R}^d$ can be split into a disjunctive union of all possible $S_{a'}$ for a given network $f(x)$:

$$\mathbb{R}^d = \cup_i \{ S_{a'_i} : \exists\, x \in \mathbb{R}^d \text{ s.t. } f(x) \text{ gives activation pattern } a'_i \} \tag{26}$$

At worst, for each possible activation pattern a' there will be an input x that produces it. It means that $\mathbb{R}^d$ will be split into at most $k \leq 2^{m \times H}$ regions, where $f(x)$ is linear on each of that regions. This number is exponentially large but finite, so it's sufficient as a proof of existence. □

Approach 2: In his lecture notes (Theorem 6.2) Matus Telgarsky takes more sophisticated approach that yields the estimate

$$k \leq \left(\frac{2m}{H}\right)^H \tag{27}$$

This is still exponentially large number, yet it's uniformly smaller than what the Approach 1 yields. It should come as no surprise, given that MJT actually counts the possible number of linear pieces as a result of superposition, instead of using a worst-case upper-bound. Nevertheless, both approaches show that the partitioning exists and the number of parts is finite. □

Problem 3

Solution for 3.1 First we notice that if a function $f(x)$ is L-positive-homogeneous then we have $f(0) = 0$. Indeed, let's take $\alpha > 0$, $\beta > 0$, $\alpha \neq \beta$. Then $\alpha^L \neq \beta^L$ for any $L \geq 0$. However, due to L-positive-homogeneity

$$\alpha^L f(0) = f(\alpha * 0) = f(0) = f(\beta * 0) = \alpha^L f(0) \tag{28}$$

which can only be true if $f(0) = 0$. This result makes it trivial to show that when $x = 0$ we get:

$$s^T x = 0 = L * 0 = Lf(0) = Lf(x) \tag{29}$$

for any s.

Solution for 3.2 Let's consider $x \neq 0$ such that $f(x)$ is differentiable in x and so $\nabla f(x)$ exists at that point. Due to the definition of the gradient we have

$$\begin{aligned} 0 &= \lim_{\delta \to 0} \frac{f(x+\delta x) - f(x) - \nabla f(x)^T \delta x}{\delta} = \\ &= \lim_{\delta \to 0} \frac{(1+\delta)^L - 1}{\delta} f(x) - \nabla f(x)^T x = \\ &= Lf(x) - \nabla f(x)^T x \end{aligned} \tag{30}$$

where we used that $(1+\delta)^L = 1 + \delta L + o(\delta)$ to calculate the limit. □

Solution for 3.3 Let $x \neq 0$ such that Clarke differential $\partial f(x)$ is defined. For every $s \in \partial f(x)$ at least one of two holds:

1. there is a sequence x_i with $\lim x_i = x$ such that $\lim \nabla f(x_i) = s$, or
2. $s = \sum_j c_j s_j$ where $\sum_j c_j = 1$ and for every s_j part (1) holds.

Let's start with the fist case. Here we have that

$$x^T s = \lim_{i \to \infty} x_i^T \nabla f(x_i) = \lim_{i \to \infty} Lf(x_i) = Lf(x) \tag{31}$$

For the second case, we get that

$$x^T s = x^T \left(\sum_j \alpha_j s_i\right) = \sum_j \alpha_j x^T s_j = \sum_j \alpha_j Lf(x) = Lf(x) \tag{32}$$

which concludes the proof. □