STAT205 HW4

June 7, 2022

Isaac K-M

Note that submissions were written up with respect to the 6/6/2022 version of the homework, as opposed to that posted the morning of 6/7/2022.

Problem 1

Note that x_i here is equivalent to $x^{(i)}$, i.e. marking a training datapoint.

a.)

Recall from HW4 that we have

$$\begin{aligned} y_i &= (\Phi\theta)_i \\ &= \theta_0 + \sum_{j=1}^{n-1} \theta_j (x_i - x_j)_+ \\ &= \theta_0 + \sum_{j=1}^{n-1} \theta_j (1 \cdot x_i - x_j)_+, \end{aligned}$$

and more generally

$$\begin{aligned} \hat{y}(x) &= \theta_0 + \sum_{j=1}^{n-1} \theta_j (x - x_j)_+ \\ &= \underbrace{\theta_0}_{\Theta_0} + \sum_{j=1}^{n-1} \underbrace{\theta_j}_{\Theta_{w_2,j}} (\underbrace{1}_{\Theta_{w_2,j}} \cdot x - \underbrace{x_j}_{\Theta_{b,j}})_+. \end{aligned}$$

This straightforwardly gives

- $\Theta_0 = \theta_0$

- $\Theta_{w_1,j} = 1$ for $j = 1, \dots n-1$
- $\Theta_{w_2,j} = \theta_j$ for $j = 1, \dots n-1$
- $\Theta_{b,j} = x_j$ for $j = 1, \dots n-1$

Hence, we can write

$$\Theta(\theta) = (\theta_0, \{x_j\}_{j=1}^{n-1}, \mathbf{1}_{n-1}, \{\theta_j\}_{j=1}^{n-1}).$$

As we see above, the relationship is 1-1, so we can similarly write the reverse via

$$\theta(\Theta)_j = \begin{cases} \Theta_0 & j = 0 \\ \Theta_{w_2,j} & j > 0 \end{cases}.$$

b.)

Recall in HW4 that the regularization scheme was given by

$$\lambda Q(\theta) = \lambda ||M\theta||_2^2 = \lambda \theta^T M^T M \theta.$$

As we showed in HW4, the entries of $M^T M$ correspond to the number of times in the training data that the relu term associated with each coefficient was activated, i.e.

$$M^T M = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & n-1 & n-2 & n-3 & n-4 & \dots & 1 \\ 0 & n-2 & n-2 & n-3 & n-4 & \dots & 1 \\ 0 & n-3 & n-3 & n-3 & n-4 & \dots & 1 \\ 0 & n-4 & n-4 & n-4 & n-4 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

However in this setting, we only have

$$\lambda Q(\theta) = \lambda ||\theta||_2^2 = \lambda \theta^T \theta,$$

i.e. there is no "weight" matrix sandwiched between the inner product of θ with itself. In this way, each θ_j^2 term in $\theta^T \theta$ contributes evenly to the penalty (this was not the case in HW4), and the solution is a proper ridge/L2 (kernel) regression.

Now, assuming we are in the same special setting where Θ has 1-1 natural correspondence with θ, with

- $\Theta_0 = \theta_0$
- $\Theta_{w_1,j} = 1$ for $j = 1, \dots n-1$
- $\Theta_{w_2,j} = \theta_j$ for $j = 1, \dots n-1$

- $\Theta_{b,j} = x_j$ for $j = 1, \ldots n-1$,

we can leverage the natural 1-1 correspondence between Θ and θ to reconstruct a ridge penalty through for Θ. Already, under this correspondence, we have

$$f(x_i; \Theta) = \phi(x_i)^T \theta,$$

so if we set

$$Q(\Theta) = \Theta_{w_2}^T \Theta_{w_2},$$

we will have on substitution

$$\begin{aligned} n^{-1} \sum_{i=1}^{n} (y_i - f(x_i; \Theta))^2 + \lambda Q(\Theta) &= n^{-1} \sum_{i=1}^{n} (y_i - \phi(x_i)^T \theta)^2 + \lambda \Theta_{w_2}^T \Theta_{w_2} \\ &= n^{-1} ||y - \Phi\theta||_2^2 + \lambda \Theta_{w_2}^T \Theta_{w_2} \\ &= n^{-1} ||y - \Phi\theta||_2^2 + \lambda \theta^T \theta \\ &= n^{-1} ||y - \Phi\theta||_2^2 + \lambda ||\theta||_2^2. \end{aligned}$$

In other words, this will make the optimization task identical, introducing a natural isomorphism. And as the NN (special case) can be reduced to ridge regression here, such an optimization is in fact representable as a quadratic optimization (see last line above), and a solution exists in closed form for $\lambda > 0$.

c.)

Note that we are assumed to be out of the special case here, and back to a full 1-NN.

For this problem, first recall the following lemma, as set forth in L13.28:
Lemma

$$min \sum_j x_j^2 + y_j^2 \text{ subject to } x_j y_j = z_j$$

for some $\{z_j\}_{j \in J}$ is achieved when

$$x_j = y_j = \sqrt{z_j}.$$

Hence,

$$\sum_j x_j^2 + y_j^2 = 2 \sum_j |z_j|$$

in such scenarios. As an intuitive example, you should think about a unit circle.

Proof

Now, for the proof, first examine $f(x, \Theta)$. We can show that this really is just linear with respect to one set of coefficients (instead of two), i.e.

$$\begin{aligned} f(x;\Theta) &= \Theta_0 + \sum_{j=1}^{d_1} \Theta_{w_2,j}(\Theta_{w_1,j} \cdot x - \Theta_{b,j})_+ \\ &= \Theta_0 + \sum_{j=1}^{d_1} \Theta_{w_2,j}\Theta_{w_1,j}(x - \Theta_{b,j}/\Theta_{w_1,j})_+ \\ &= \tilde{\theta}_0 + \sum_{j=1}^{d_1} \tilde{\theta}_j(x - \tilde{b}_j)_+ \\ &= \tilde{\phi}_{\tilde{b}}(x)^T\tilde{\theta}, \end{aligned}$$

where

$$\begin{aligned} \tilde{\phi}_{\tilde{b}}(x)^T &= [1, (x - \tilde{b}_1)_+, (x - \tilde{b}_2)_+, \ldots, (x - \tilde{b}_{d_1})_+] \\ &= \left[1, \left(x - \frac{\Theta_{b,1}}{\Theta_{w_1,1}}\right)_+, \left(x - \frac{\Theta_{b,2}}{\Theta_{w_1,2}}\right)_+, \ldots, \left(x - \frac{\Theta_{b,d_1}}{\Theta_{w_1,d_1}}\right)_+\right] \end{aligned}$$

akin to HW4. This is the linear prediction with which we're familiar, albeit with new knots $\{\tilde{b}_j\}_{j=1}^{d_1}$ and weights $\{\tilde{\theta}_j\}_{j=0}^{d_1}$. That is, the dual-weighted Θ setup has an isomorphism with a single-weighted linear setup.

We can then vectorize the MSE portion of the objective to

$$\begin{aligned} \sum_{i=1}^{N}(y_i - f(x_i;\Theta))^2 &= \sum_{i=1}^{N}(y_i - \tilde{\phi}_{\tilde{b}}(x_i)^T\tilde{\theta})^2 \\ &= ||y - \Phi_\Theta\tilde{\theta}||_2^2, \end{aligned}$$

where (again, we'll keep with the zero-indexing of columns and coefficients) the basis is

$$\Phi_{\Theta i,j} = \begin{cases} 1 & j = 0 \\ \left(x_i - \frac{\Theta_{b,j}}{\Theta_{w_1,j}}\right)_+ & j = 1, \ldots, d_1 \end{cases}$$

Thus, the takeaway is that minimization of

$$\sum_{i=1}^{N}(y_i - f(x_i;\Theta))^2$$

is identical to the minimization of

$$||y - \Phi_\Theta\tilde{\theta}||_2^2,$$

where Φ_Θ has been defined in terms of Θ and the training dataset $\{x_i\}_{i=1}^n$ above.

Now, suppose we have found our Θ_{opt}, and suppose that this Θ_{opt} achieves an MSE of

$$\mathcal{M}(x, \Theta_{opt}) = \sum_{i=1}^{n} (y_i - f(x_i; \Theta_{opt}))^2.$$

By our work above, we have

$$\begin{aligned}\mathcal{M}(x, \Theta_{opt}) &= n^{-1} \sum_{i=1}^{n} (y_i - f(x_i; \Theta_{opt}))^2 \\ &= n^{-1} ||y - \Phi_{\Theta_{opt}} \tilde{\theta}_{opt}||_2^2.\end{aligned}$$

Then (knowing the answer ahead of time), we can rewrite the full minimization

$$\min_{\Theta} \mathcal{M}(x, \Theta) + \lambda Q(\Theta)$$

as

$$\min_{\Theta} \lambda(||\Theta_{w_1}||_2^2 + ||\Theta_{w_2}||_2^2) = \min_{\Theta} \lambda(\sum_{j=1}^{d_1} \Theta_{w_1,j}^2 + \Theta_{w_2,j}^2)$$

subject to $\mathcal{M}(x, \Theta) = \mathcal{M}(x, \Theta_{opt})$. As shown in the relu simplification above, we have

$$\tilde{\theta}_j = \Theta_{w_1,j} \Theta_{w_2,j};$$

hence, we invoke the lemma, and see that this minimization is the same as

$$\min_{\tilde{\theta}} \lambda(\sum_{j=1}^{d_1} |\tilde{\theta}_j|) = \lambda ||\tilde{\theta}||_1.$$

subject to $\mathcal{M}(x, \Theta) = \mathcal{M}(x, \Theta_{opt})$. However, we know from above that

$$\mathcal{M}(x, \Theta) = \mathcal{M}(x, \Theta_{opt}) \implies n^{-1} ||y - \Phi_\Theta \tilde{\theta}||_2^2 = \mathcal{M}(x, \Theta_{opt}).$$

Thus, in all, we see that the Θ_{opt} that satisfies

$$\min_{\Theta} \left[\mathcal{M}(x, \Theta) + \lambda(||\Theta_{w_1}||_2^2 + ||\Theta_{w_2}||_2^2) \right]$$

corresponds directly to the $\tilde{\theta}_{opt}$ that satisfies

$$\min_{\tilde{\theta}} \left[n^{-1} ||y - \Phi_{\Theta_{opt}} \tilde{\theta}||_2^2 + \lambda ||\tilde{\theta}||_1 \right],$$

where again we have required $n^{-1}||y - \Phi_{\Theta_{opt}} \tilde{\theta}||_2^2 = \mathcal{M}(x; \Theta_{opt})$. In other words, it can be reconfigured as a Lasso. And since the Lasso can be solved via quadratic programming methods, we have a quadratic optimization over linear space. Lastly, I would consider the correspondence here to be non-linear, as a critical step involves computing a $\Phi_{\Theta_{opt}}$ via repeated applications of the ReLU, which is the most well-known *non-linear* activation function. In other words, construction of the design matrix to extract $\tilde{\theta}$ requires non-linear activity, so I would consider the correspondence non-linear.

Problem 2

a.)

Simple addition, subtraction, and multiplication properties give

$$\phi_{b_2}(\phi_{b_1}(x)) = \begin{cases} relu(x - (b_1 + b_2)) & b_2 \geq 0 \\ relu(x - b_1) & b_2 < 0. \end{cases}$$

An intuitive way to think of this is that if $b_2 > 0$, it imposes more stringent standards on the ReLU; otherwise, it adds no standards, so the standards imposed by b_1 suffice.

b.)

Here, it is helpful to split into cases. First, consider what happens when x survives the first activation, i.e. $x > b_1$. We have

$$x > b_1 \implies w_1 \cdot \underbrace{relu(x - b_1)}_{>0} = w_1(x - b_1).$$

In order to satisfy the second activation, we then need

$$w_1(x - b_1) - b_2 > 0 \implies x > b_1 + \frac{b_2}{w_1}.$$

If this is satisfied, all activations are survived; hence, we have a composition of linear functions and

$$\phi_{b_3}(x) = w_2(w_1(x - b_1) - b_2).$$

and zero otherwise.

Second, consider what happens when $x < b_1$, i.e. x does not survive the first activation. This will give

$$w_1(x - b_1) = 0,$$

so the rest of the function rests solely on b_2. If $b_2 \leq 0$, then

$$w_1(x - b_1) - b_2 = 0 - b_2 > 0$$

so the second relu will activate; otherwise, when $b_2 \geq 0$ the second relu will not activate, and return zero.

At this point, we have solved the survival logic for the activations. If we combine everything in indicators, in all we have

$$\begin{aligned} \phi_{b_3}(x) = \mathbf{1}(x > b_1)\mathbf{1}(x > b_1 + b_2/w_1)[w_2(w_1(x - b_1) - b_2)] \\ + \mathbf{1}(x \leq b_1)\mathbf{1}(b_2 < 0)[-w_2 \cdot b_2]. \end{aligned}$$

c.)

Note that the result above is derived without consideration for the sign of w_1, w_2, so the expression holds.

d.)

f_ℓ Non-decreasing
First, consider $x' > x$. For any $b_i^{(\ell)}$, it is necessarily the case that

$$relu(x' - b_i^{(\ell)}) \geq relu(x - b_i^{(\ell)}).$$

Then, since $c_i^{(\ell)} \geq 0$, we have

$$c_i^{(\ell)} \cdot relu(x' - b_i^{(\ell)}) \geq c_i^{(\ell)} \cdot relu(x - b_i^{(\ell)})$$

and hence

$$\sum_i c_i^{(\ell)} \cdot relu(x' - b_i^{(\ell)}) \geq \sum_i c_i^{(\ell)} \cdot relu(x - b_i^{(\ell)}),$$

as desired.

f_ℓ Increasing
First, let $\tilde{i}$ be the i such that b_i^ℓ satisfies $\min\{b_i^\ell : c_i^\ell > 0\}$.

By construction, b_i^ℓ is the term (for which the coefficient is nonzero) which spends the least time at zero; i.e. it's the most "easily" activated. Hence, for any $x > x_{0,\ell}$, we have

$$relu(x - x_{0,\ell}) = relu(x - b_{\tilde{i}}^\ell) > 0 \implies c_{\tilde{i}}^\ell \cdot relu(x - x_{0,\ell}) = c_{\tilde{i}}^\ell \cdot relu(x - b_{\tilde{i}}^\ell) > 0$$

Then, for any $x' > x$, we similarly have (just as before, since $x' > x_{0,\ell}$)

$$c_{\tilde{i}}^\ell \cdot relu(x' - x_{0,\ell}) = c_{\tilde{i}}^\ell \cdot relu(x' - b_{\tilde{i}}^\ell) > 0$$

And since we have: (i) $x' > x$; (ii) $c_{\tilde{i}}^\ell \cdot relu(x - x_{0,\ell}) > 0$; and (iii) $c_{\tilde{i}}^\ell \cdot relu(x' - x_{0,\ell}) > 0$, we are guaranteed

$$c_{\tilde{i}}^\ell \cdot relu(x' - x_{0,\ell}) > c_{\tilde{i}}^\ell \cdot relu(x - x_{0,\ell}),$$

or

$$c_{\tilde{i}}^\ell \cdot relu(x' - b_{\tilde{i}}^\ell) > c_{\tilde{i}}^\ell \cdot relu(x - b_{\tilde{i}}^\ell).$$

From before, we know that for all other $i \neq \tilde{i}$

$$c_i^\ell \cdot relu(x' - b_i^\ell) > c_i^\ell \cdot relu(x - b_i^\ell).$$

This then gives

$$\begin{aligned}\sum_{i=1}^{N} c_i^\ell \cdot relu(x' - b_i^\ell) &= c_{\tilde{i}}^\ell \cdot relu(x' - b_{\tilde{i}}^\ell) + \sum_{i \neq \tilde{i}} c_i^\ell \cdot relu(x' - b_i^\ell) \\ &> c_{\tilde{i}}^\ell \cdot relu(x - b_{\tilde{i}}^\ell) + \sum_{i \neq \tilde{i}} c_i^\ell \cdot relu(x' - b_i^\ell) \\ &\geq c_{\tilde{i}}^\ell \cdot relu(x - b_{\tilde{i}}^\ell) + \sum_{i \neq \tilde{i}} c_i^\ell \cdot relu(x - b_i^\ell) \\ &= \sum_{i=1}^{N} c_i^\ell \cdot relu(x - b_i^\ell),\end{aligned}$$

showing the desired increase.

f_3 Non-decreasing
First, observe

$$\begin{aligned}f_3(x) &= f_2(f_1(x)) \\ &= \sum_{i=1}^{n_2} c_i^2 \phi_{b_i^2}\left(\sum_{j=1}^{n_1} c_j^1 \phi_{b_i^1}(x)\right)\end{aligned}$$

Now, suppose that $x' > x$, but that $f_3(x') < f_3(x)$. By the construction above, this necessarily implies that there exists some i^* such that

$$c_i^{*,2} \phi_{b_i^{*,2}}\left(\sum_{j=1}^{n_1} c_j^1 \phi_{b_i^1}(x)\right) < 0.$$

However, since $c_i^{*,2} > 0$ and since the $\phi_{b_i^{*,2}} : \mathbb{R} \to \mathbb{R}^+ \cup \{0\}$, we necessarily have a contradiction – that is, it is impossible for this term to be sub-zero. This gives non-decreasing.

f_3 Increasing

First, if it is the case that $x_{0,3} = x_{0,1}$, then increasingness on $x > x_{0,3} = x_{0,1}$ follows from the increasingness proof above: we know that $f_1 = c_i^1 \phi_{b_i^1}(x)$ is increasing on this domain. Then, since f_2 is non-decreasing, we have over $x > x_{0,3} = x_{0,1}$ a non-decreasing function of an increasing function, which is increasing.

Second, if it is the case that $x_{0,3} = f_1^{-1}(x_{0,2}$, then we have (reintroducing our $\tilde{i}$ notation

for the bias that satisfies the minimum).

$$\begin{aligned} f_1^{-1}(x_{0,2}) &= \sup\{x : f_1(x) \leq x_{0,2}\} \\ &= \sup\{x : \sum_{i=1}^{n_1} c_1^1 \cdot relu(x - b_i^1) \leq x_{0,2}\} \\ &= \sup\{x : \sum_{i=1}^{n_1} c_1^1 \cdot relu(x - b_i^1) \leq \min\{b_i^2 : c_i^2 > 0\}\} \\ &= \sup\{x : \sum_{i=1}^{n_1} c_1^1 \cdot relu(x - b_i^1) \leq b_{\tilde{i}}^2\}. \end{aligned}$$

The supremum will be satisfied at equality, and hence we will have

$$x_{0,3} \quad s.t. \quad \sum_{i=1}^{n_1} c_1^1 \cdot relu(x_{0,3} - b_i^1) = f_1(x_{0,3}) = b_{\tilde{i}}^2.$$

Now, take any $x' > x \geq x_{0,3}$. Then, because we are in case two and $x_{0,3} > x_{0,1}$, we have

$$f_1(x') > f_1(x) > f_1(x_{0,3}) = b_{\tilde{i}}^2.$$

For the $\tilde{i}$ term, we have

$$c_{\tilde{i}}^2 \cdot relu(f_1(x_{0,3}) - b_{\tilde{i}}^2) = 0$$

by construction, but

$$c_{\tilde{i}}^2 \cdot relu(f_1(x) - b_{\tilde{i}}^2) > 0$$

and

$$c_{\tilde{i}}^2 \cdot relu(f_1(x') - b_{\tilde{i}}^2) > 0$$

because

$$f_1(x') > f_1(x) > f_1(x_{0,3}) = b_{\tilde{i}}^2.$$

And since $f_1(x') > f_1(x)$, we get

$$c_{\tilde{i}}^2 \cdot relu(f_1(x') - b_{\tilde{i}}^2) > c_{\tilde{i}}^2 \cdot relu(f_1(x) - b_{\tilde{i}}^2).$$

Then, for all other $i \neq \tilde{i}$, we have

$$\sum_{i \neq \tilde{i}} c_i^2 \cdot relu(f_1(x') - b_i^2) \geq \sum_{i \neq \tilde{i}} c_i^2 \cdot relu(f_1(x) - b_i^2).$$

Thus, putting everything together, we get

$$\begin{aligned}
f_2(f_1(x')) &= \sum_{i=1}^{n_2} c_i^2 \cdot relu(f_1(x') - b_i^2) \\
&= c_{\tilde{i}}^2 \cdot relu(f_1(x') - b_{\tilde{i}}^2) + \sum_{i \neq \tilde{i}} c_i^2 \cdot relu(f_1(x') - b_i^2) \\
&> c_{\tilde{i}}^2 \cdot relu(f_1(x) - b_{\tilde{i}}^2) + \sum_{i \neq \tilde{i}} c_i^2 \cdot relu(f_1(x') - b_i^2) \\
&\geq c_{\tilde{i}}^2 \cdot relu(f_1(x) - b_{\tilde{i}}^2) + \sum_{i \neq \tilde{i}} c_i^2 \cdot relu(f_1(x) - b_i^2) \\
&= \sum_{i=1}^{n_2} c_i^2 \cdot relu(f_1(x') - b_i^2) \\
&= f_2(f_1(x)),
\end{aligned}$$

as desired. Again, we have contained ourself to $x > x_{0,3}$ in order to introduce keep ourselves on the active area of the ReLU, in order for this to work.

f_3 Piece-wise Linear
At this point in the class, the following two facts should be clear:

1.) The composition of piece-wise linear function f (N_f unique nontrivial knots) with a piece-wise linear function g (N_g unique nontrivial knots) will itself be a piece-wise linear function, with as many as $N_g + N_f$ knots. If knots align perfectly, then this composition will have $N_{fg} = N_g = N_f$ knots (i.e. the same amount); if there are no shared knots across f, g, then the new composition will have $N_{fg} = N_g + N_f$. So under composition, you get a new piece-wise linear function, with as few as $\max\{N_g, N_f\}$ new knots and as many as $N_g + N_f$ new knots. Note that some of these new knots may be trivial (i.e. slope approaching from both sides is the same), depending on the composition going on on the LHS and RHS of the knot in question.

To see the mechanics of this consider the following, for $f \circ g$, where f and g are piece-wise linear with knot sets b_f, b_g. First, concatenate and sort (ascending) $\tilde{b} = sort(\{b_f, b_g\})$ to get a "new" knot set. Second, examine each $\tilde{b}_i, \tilde{b}_{i+1}$ (a "chunk"). Within this chunk, for $x \in [\tilde{b}_i, \tilde{b}_{i+1}]$, $f \circ g$ on x will be linear, since a composition of linear functions is linear. Do this over every "chunk", and each chunk will be linear; hence, a piece-wise linear function in all.

2.) The addition of piece-wise linear functions is also a piece-wise linear function. The logic/intuition is identical to that above – however now, within each chunk, we add the linear functions instead of compose them (i.e. multiply).

These two facts give us everything we need to iteratively show piece-wise linearity. Already, we know that

$$c_i^1 \cdot relu(x - b_i^1)$$

is piece-wise linear (last lecture; basic intuition). By the additive property, it follows that

$$f_1(x) \sum_i c_i^1 \cdot relu(x - b_i^1)$$

must also be piece-wise linear. Similarly, we know that

$$c_j^2 \cdot relu(z - b_j^2)$$

is piece-wise linear in z. Since a composition of piece-wise linear functions piece-wise linear, we know that

$$c_j^2 \cdot relu(f_1(x) - b_j^2)$$

is also piece-wise linear too. Lastly, using the additive property again, we have that

$$\sum_j c_j^2 \cdot relu(f_1(x) - b_j^2) = f_2(f_1(x))$$

must also be piece-wise linear, completing the proof.

Where f_3 Has Knots

Since each $relu(x - b_i^\ell)$ is a piecewise function with a knot at b_i^ℓ, the sum

$$f_1(x) = \sum_{i=1}^{n_1} c_i^1 \cdot relu(x - b_1^\ell)$$

has knots/kinks at $\{b_i^1\}$. Intuitively, any time as x crosses over b_i^1 (i.e. goes from $x < b_i^1$ to $x > b_i^1$), the function

$$\sum_{k \neq i} c_k^1 \cdot relu(x - b_k^1)$$

may be smooth in x, but the non-linear activation will still exist for the $c_i^1 \cdot relu(x - b_1^1)$, so overall the function will still have a knot there. This happens over all $i = 1, \ldots, n_1$ biases; hence a guaranteed $\{b_i^1\}_{i=1}^{n_1}$ knots. No matter how else you compose f_1 with some other f_2, these initial knots – as they directly interact with the x on entry to the model – will remain. Intuitively, they'll be in on the ground floor.

Next, you may also see knots at $\{f_1^{-1}(b_j^2)$. If we expand out this term a bit, the intuition becomes clear

$$\{f_1^{-1}(b_j^2) = \sup\{x : f_1(x) \leq b_j^2\},$$

and we know that for $x : f_1(x) > b_j^2$, we will have activation

$$c_j^2 \cdot (f_1(x) - b_j^2) \neq 0,$$

whereas for $x : f_1(x) \leq b_j^2$, we will have no activation

$$c_j^2 \cdot (f_1(x) - b_j^2) = 0,$$

In other words, x crosses some threshold that, through f_1 switches on an activation two layers ahead. In this way, it introduces a piece-wise linearity (via the relu) in $f_1(x)$ at b_j^1; hence, if we take that threshold with respect to x instead of $f_1(x)$, the threshold is $f_1^{-1}(b_j^1)$.

Of course, not all of these points will come to pass as additional knots. For instance, $f_1(x)$ maybe configured such that activation of

$$relu(f_1(x) - b_2^j)$$

is impossible, in which case no knot would be added. So it is possible, but not certain, to add knots in this manner.

With the knot sets for f_2 and f_1 established above, it is clear that the knot set for their composition, f_3 is

$$\{b_i^1\} \cup \{f_1^{-1}(b_j^2)\}.$$

The first set $\{b_i^1\}$ follows from the first commentary above; the second set $\{f_1^{-1}(b_j^2)\}$ follows from the second half, where x induces a threshold crossing/activation in the second layer, and hence a knot. In the event that such a threshold crossing is induced in all n_2 functions in the second layer, this will tack on an additional n_2 knots, giving $n_1 + n_2$ knots in all.

Conclude/Explain
Our results above incrementally and logically lead us to the stated conclusion. The first part of the sentence is given by the shown piecewise linearity above; the second part of the sentence is given by the N_g, N_f discussion in the piecewise linearity proof. That is, when you compose or add a linear piecewise function with N_f knots with a linear piecewise function with N_g knots, there are $N_f + N_g$ "chunks" to evaluate the composition over, occurring at the combined set of all knots. This leaves ample room for $N_f + N_g$ total knots; however, it is likely that the combined piece-wise linearities – either through the relu, multiplication, or addition, that some of these knots will be trivialized (e.g. LHS slope = RHS slope), so there may well be fewer knots.

e.)

See Colab pdf.

3.)

See Colab pdf.