Gatsby Computational Neuroscience Unit
Theoretical Neuroscience

Final examination, theoretical neuroscience
6-8 May 2021

Part II – long questions

There are four questions, one from each main section of the course. Please answer three out of the four, starting the answers for each question on a new page. Don't forget to write your name at the top of the answer to each question.

Good luck!

1 Biophysics

Consider a linear integrate and fire neuron whose voltage, V, evolves according to

$$\tau \frac{dV}{dt} = -(V - \mathcal{E}_L) + V_0 - g(V - \mathcal{E}_K) \tag{1}$$

where $\mathcal{E}_L = -65$ mV and $\mathcal{E}_K = -80$ mV, and V_0 is a free parameter. When the neuron reaches threshold (taken to be -50 mV), it emits a spike and is reset to $\mathcal{E}_L$. The potassium conductance, g, increases whenever the neuron fires (but with a delay), and then decays back to zero,

$$\frac{dg}{dt} = \Delta g \sum_k \delta(t - t_d - t_k) - \frac{g}{\tau_K} \tag{2}$$

where t_k is the time of the k^{th} spike, τ_d is the delay, and Δg is a free parameter.

As usual, we want to get rid of as many free parameters as possible, and so we define $u \equiv V + \mathcal{E}_L$, and work with u rather then V,

$$\tau \frac{du}{dt} = -u + V_0 - g(u + \Delta) \tag{3} \quad \{\texttt{lif}\}$$

where $\Delta = \mathcal{E}_L - \mathcal{E}_K = 15$ mV, and the neurons spikes when $u = \theta = 15$ mV.

1. Show that when g is fixed, the time, $t_s(g)$, it takes the neuron to emit a spike (starting from $u = 0$) is

$$t_s(g) = \frac{\tau}{1+g} \log \frac{V_0 - g\Delta}{V_0 - g\Delta - \theta(1+g)} \,. \tag{4} \quad \{\texttt{ts}\}$$

   (4 marks)

2. Show that the neuron stops firing when

$$g \geq g^* \equiv \frac{V_0 - \theta}{\Delta + \theta} \,. \tag{5} \quad \{\texttt{gs}\}$$

   (3 marks).

3. Set $\tau_K = \infty$ and $t_d = 0$, and assume $V_0 > \theta$. If initially $g = 0$, show that the neuron will emit n_{spikes} spikes and then stop firing, where

$$n_{\text{spikes}} = \left\lceil \frac{1}{\Delta g} \frac{V_0 - \theta}{\Delta + \theta} \right\rceil \,. \tag{6}$$

   Here $\lceil x \rceil$ denotes the smallest integer greater than or equal to x.

   (3 marks)

4. Assume $t_d \gg \tau_K \gg t_s(0)$ (see problem 1 for the last quantity, $t_s(g)$), and $t_d/t_s(0) > n_{\text{spikes}}$. Describe the behavior of the neuron (that is, describe the spike trains).

   (15 marks)

5. Assume there is no delay ($t_d = 0$), and $\tau_K \gg \tau$ (the time constant of the neuron). Show that, to good approximation, the neuron fires regularly and the time between spikes, t^*, is

$$t^* = \tau_k \log \frac{g^* + \Delta g}{g^*} \tag{7}$$ {tnd}

   where g^* is given in Eq. (5).

   (15 marks)

Solutions

1. We can rewrite Eq. (3) as

$$\frac{\tau}{1+g}\frac{du}{dt} = -u + \frac{V_0 - g\Delta}{1+g}\,. \tag{8}$$

   When g is constant and $u(t=0) = 0$, this has the solution

$$u(t) = \frac{V_0 - g\Delta}{1+g}\left(1 - e^{-t(1+g)/\tau}\right)\,. \tag{9}$$

   A spike occurs when $u(t) = \theta$. Setting the left hand side to θ thus gives us $t_s(g)$, the time of a spike. A small amount of algebra gives us Eq. (4).

2. From Eq. (4), we see that the neuron fires so long as

$$V_0 - g\Delta > \theta(1+g)\,. \tag{10}$$

   Solving for g yields Eq. (5).

3. If $\tau_K = \infty$, then every time there's a spike, g increases by Δg. The neuron stops spiking when $g > (V_0 - \theta)/(\Delta + \theta)$ (since there's no delay). The number of spikes it takes for this to occur is

$$\frac{1}{\Delta g}\,\frac{V_0 - \theta}{\Delta + \theta}\,, \tag{11}$$

   rounded up to the nearest integer.

4. We'll assume that we're starting from $u = 0$ and $g = 0$. The neuron emits $t_d/t_s(0)$ spikes before g starts to increase, and after time $2t_d$, all those spikes will have affected g. Because $\tau_K \gg t_s(0)$, g will increase by about $\Delta g t_d/t_s(0)$, and because $t_d/t_s(0) > n_{\text{spikes}}$ this is enough to stop the neuron from firing. When that happens, g will decay with time constant τ_K. Eventually it will get small enough for the neuron to fire, but that won't affect g for a time t_d. Since $t_d \gg \tau_K$, by the time the conductance starts to rise again, u will be near zero. At that point, the cycle will start over again.

   To summarize, the neuron "bursts": it's silent for a time that scales with t_d, and then fires for a time that scales with t_d.

5. Let's start, as usual, with $g = 0$. The neuron will fire until $g > g^*$, at which point it will stop firing. If $\tau_K \gg \tau$, then we can treat g as nearly constant, and so the membrane potential will quickly equilibrate, at some value below threshold, yielding (via Eq. (3)),

$$u(t) = \frac{V_0 - g(t)\Delta}{1 + g(t)} \tag{12}$$

where $g(t) \propto e^{-t/\tau_K}$. When $g(t)$ decays to g^*, the neuron will spike, and g will jump to $g^* + \Delta g$. It will then decay again, via

$$g(t) = (g^* + \Delta g)e^{-t/\tau_K} \tag{13}$$

where t is time relative to a spike. To find the time of the next spike, t^*, we need to solve the equation

$$(g^* + \Delta g)e^{-t/\tau_K} = g^* \,. \tag{14}$$

Solving for t^* yields Eq. (4).

2 Networks

Consider a network in which the activity is coupled via a low rank matrix,

$$\frac{dx_i}{dt} = \phi\left(\sum_{j=1}^{n} J_{ij}x_j\right) - x_i \tag{1}$$ {xdot}

where

$$J_{ij} \equiv \frac{1}{n}\sum_{\mu=1}^{p} \eta_{\mu,i}\xi_{\mu,j} \tag{2}$$ {J}

with $p \ll n$. The elements of η and ξ are drawn from a zero mean Gaussian distribution with the following covariance structure, {CC}

$$\langle \eta_{\mu,i}\eta_{\nu,i}\rangle = \delta_{\mu\nu} \tag{3a}$$
$$\langle \xi_{\mu,i}\xi_{\nu,i}\rangle = \delta_{\mu\nu} \tag{3b}$$
$$\langle \xi_{\mu,i}\eta_{\nu,i}\rangle = C_{\mu\nu} \tag{3c}$$

where $\delta_{\mu\nu}$ is the Kronecker delta. In words, the $\eta_{\mu,i}$ are uncorrelated with each other, the $\xi_{\mu,i}$ are uncorrelated with each other, and the $\xi_{\mu,i}$ are correlated with the $\eta_{\mu,i}$, with correlation coefficient $C_{\mu\nu}$.

1. Define

$$m_\mu \equiv \frac{1}{n}\sum_j \xi_{\mu,j}x_j\,. \tag{4}$$

Show that in the large n limit, m_μ evolves according to

$$\frac{dm_\mu}{dt} = \left\langle \xi_\mu \phi\left(\sum_\nu m_\nu\eta_\nu\right)\right\rangle_{\xi_\mu,\boldsymbol{\eta}} - m_\mu\,. \tag{5}$$ {mdot}

The variables ξ_μ and $\boldsymbol{\eta}$ are Gaussian distributed, with the covariance structure given in Eq. (3) (but without the subscript i).

(10 marks)

2. To perform the Gaussian averages, note that the argument of ϕ obeys {averages}

$$\sum_\nu m_\nu\eta_\nu \sim \mathcal{N}\left(0, \mathbf{m}\cdot\mathbf{m}\right) \tag{6a}$$

$$\left\langle \xi_\mu \sum_\nu m_\nu\eta_\nu\right\rangle = \sum_\nu C_{\mu\nu}m_\nu \tag{6b}$$

where $\mathbf{m}\cdot\mathbf{m} = \sum_\nu m_\nu^2$. Combining this with the fact that, for any differentiable function $f(z)$,

$$\int dz \frac{e^{-z^2/2}}{\sqrt{2\pi}}\, z f(z) = \int dz \frac{e^{-z^2/2}}{\sqrt{2\pi}}\, f'(z) \tag{7}$$ {gaussint}

where prime denotes a derivative (which is not hard to show), show that

$$\frac{dm_\mu}{dt} = f(|\mathbf{m}|)\sum_\nu C_{\mu\nu} m_\nu - m_\mu \tag{8}$$ {mdot2}

where $|\mathbf{m}| \equiv \sqrt{\mathbf{m}\cdot\mathbf{m}}$ and

$$f(|\mathbf{m}|) \equiv \int dz \frac{e^{-z^2/2}}{\sqrt{2\pi}}\, \phi'(|\mathbf{m}|z)\,, \tag{9}$$

In vector notation, this becomes

$$\frac{d\mathbf{m}}{dt} = f(|\mathbf{m}|)\,\mathbf{C}\cdot\mathbf{m} - \mathbf{m} \tag{10}$$ {m}

(10 marks)

3. The fixed point at $\mathbf{m} = 0$ is stable. Our question is: are there other fixed points? If they are, they must be proportional to the eigenvectors of $\mathbf{C}$. There are p such eigenvectors (recall that the $\mathbf{C}$ is $p \times p$; see Eq. (2) above). Give conditions for the existence of these p solutions

   (5 marks)

4. Show that only the solution corresponding to the eigenvector with largest eigenvalue can be stable (i.e., the solutions *not* corresponding to the largest eigenvalue are unstable).

   (15 marks)

Solutions

1. Note first of all that

$$\sum_j J_{ij} x_j = \sum_\mu \eta_{\mu,i} \frac{1}{n} \sum_j \xi_{\mu,j} x_j = \sum_\mu m_\mu \eta_{\mu,i}\,. \tag{11}$$

   Using this, and multiplying both sides of Eq. (1) with $\xi_{\mu,i}/n$ and summing over i yields

$$\frac{dm_\mu}{dt} = \frac{1}{n}\sum_i \xi_{\mu,i}\phi\left(\sum_\nu m_\nu \eta_{\nu_i}\right) - m_\mu\,. \tag{12}$$

   Finally, if we replace the sum over i by an average over the ξ and η, we recover Eq. (5).

2. Given Eq. (6), we can write

$$\left\langle \xi_\mu \phi \left(\sum_\nu m_\nu \eta_\nu \right) \right\rangle_{\xi_\mu, \boldsymbol{\eta}} = \langle \xi_\mu \phi(|\mathbf{m}|z) \rangle_{\xi_\mu, z} \tag{13}$$

where $z \sim \mathcal{N}(0,1)$ and

$$\langle \xi_\mu z \rangle = \frac{1}{|\mathbf{m}|} \sum_\nu C_{\mu\nu} m_\nu \,. \tag{14}$$ {rho}

Let us make the change of variables

$$\eta_\mu = \eta'_\mu + \frac{z}{|\mathbf{m}|} \sum_\nu C_{\mu\nu} m_\nu \,. \tag{15}$$

To satisfy Eq. (**??**), we must have $\langle \eta'_\mu z \rangle = 0$. Consequently,

$$\begin{aligned} \langle \eta_\mu \phi(|\mathbf{m}|z) \rangle_{\eta_\mu, z} &= \langle \eta'_\mu \phi(|\mathbf{m}|z) \rangle_{\eta'_\mu, z} + \frac{1}{|\mathbf{m}|} \sum_\nu C_{\mu\nu} m_\nu \langle z \phi(|\mathbf{m}|z) \rangle_z \\ &= \frac{1}{|\mathbf{m}|} \sum_\nu C_{\mu\nu} m_\nu \langle z \phi(|\mathbf{m}|z) \rangle_z \,. \end{aligned} \tag{16}$$

For the remaining term on the right, we use Eq. (7), which leads, after a small amount of algebra, to Eq. (8)

3. Let $\mathbf{v}_k$ be the k^{th} (orthonormal) eigenvector of $\mathbf{C}$ and λ_k be the corresponding eigenvalue. Let $\mathbf{m} = \alpha_k \mathbf{v}_k$. If this is an equilibrium, it must obey

$$f(\alpha_k)\lambda_k = 1 \,. \tag{17}$$

If this equation can be satisfied (which it typically can be), then $\alpha_k \mathbf{v}_k$ is an equilibrium.

4. To check or stability, we linearize by letting

$$\mathbf{m} = \alpha_k \mathbf{v}_k + \delta\mathbf{m} \tag{18}$$

where $\alpha_k \mathbf{v}_k$ is a fixed point. Linearizing gives

$$\frac{d\delta\mathbf{m}}{dt} = 2\alpha_k f'(\alpha_k) \mathbf{v}_k \cdot \delta\mathbf{m} + f(\alpha_k) \mathbf{C} \cdot \delta\mathbf{m} - \delta\mathbf{m} \,. \tag{19}$$

Suppose λ_k is not the largest eigenvalue. Let $\delta\mathbf{m} = \alpha_l(t)\mathbf{v}_l$ with $\lambda_l > \lambda_k$. Because $\mathbf{v}_k \cdot \mathbf{v}_l = 0$, we have

$$\frac{d\alpha_l}{dt} = \big(f(\alpha_k)\lambda_l - 1\big)\alpha_l \,. \tag{20}$$

The equilibrium condition for the fixed point $\alpha_k \mathbf{v}_k$ is

$$f(\alpha_k)\lambda_k = 1 \,. \tag{21}$$

Consequently, we can write the above equation as

$$\frac{d\alpha_l}{dt} = \left(\frac{\lambda_l}{\lambda_k} - 1 \right) \alpha_l \,. \tag{22}$$

Because $\lambda_l > \lambda_k$, the term in parentheses is positive, and the equilibrium is unstable.

3 Coding

The joint peri-stimulus-time histogram (JPSTH) is a representation of the joint activity of two cells in the form of a two-dimensional array of bins. Let $t_i^{(c,k)}$ represent the time of the ith spike from cell c during trial k. Then the (m,n)th bin of the JPSTH counts the average number of times a spike from cell 1 falls in the mth PSTH bin and a spike from cell 2 falls in the nth PSTH bin (of width Δ) on the same trial; that is, on each trial k we count the number of pairs (i,j) such that $t_i^{(1,k)} \in [(m-1)\Delta, m\Delta)$ and $t_j^{(2,k)} \in [(n-1)\Delta, n\Delta)$ and then average over trials.

Consider two cells which both respond to a stimulus by firing with the identical mean firing rate $\overline{\lambda}(t) = \rho(t)$.

1. If both cells fire according to (independent) inhomogeneous Poisson processes, what is the expected value of the JPSTH? (You can assume that the mean rates vary slowly enough to treat $\rho(t)$ as constant within a bin). (2 marks)

2. Suppose now that each cell fires according to a doubly-stochastic inhomogeneous Poisson process, whose intensity on the kth trial is given by $\lambda^{(k)}(t) = g^{(k)}\rho(t)$. Here, $g^{(k)}$ is a gain or scale that varies randomly from trial-to-trial with mean 1 and variance γ^2. Both cells are scaled by the same factor on the same trial. Now what is the expected value of the JPSTH? (7 marks)

3. An alternative form of correlation might be due to common drive from another cell. Suppose that a third, unobserved, cell fires spikes according to a Poisson process with a rate $\rho^{(3)}(t)$. Let $s^{(3,k)}(t)$ be a spike train from this cell, written as a sum of delta-functions. Then we assume that the observed cells each fire with point process intensities $\lambda^{(c,k)}(t) = (1-\alpha)\rho(t) + \alpha \int d\tau\ k(t-\tau)s^{(3,k)}(\tau)$. Take k to be a causal positive filter and assume that $\int d\tau\ k(t-\tau)\rho^{(3)}(\tau) = \rho(t)$. What is the expected JPSTH in this case? (You can leave your answer in the form of a single—but not multiple—integral). (13 marks)

In principle, we can construct models that predict the JPSTH (or the corresponding trial-by-trial joint counts) from a stimulus as we do for single cell firing.

Specifically, consider a general LN model, which looks for a "joint" stimulus subspace from which the simultaneous responses of the two cells can be predicted. That is, we project a high-dimensional stimulus vector at time t, $\mathbf{x}_t$ onto one or more vectors $\mathbf{k}_d, d = 1 \ldots D$, with the expected values of the JPSTH cells depending nonlinearly on these projections through some unknown functions. For a single spike-train, we would associate the vector $\mathbf{x}_t$ with the spike count at time t. In the JPSTH, the joint counts collect spikes from two different times in the two cells, say t and t'. Let $t' > t$. We associate with the joint count the stimulus $\mathbf{x}_{t'}$, and assume that this vector looks far enough back in time to be reasonably able to capture the behaviour at time t as well.

Thus, we construct a weighted spike-triggered ensemble by associating with $\mathbf{x}_t$ a weight which corresponds to the sum of counts in all the JPSTH bins with which the stimulus is associated (just as in single-cell spike-triggered analysis we can associate each stimulus with the number of spikes generated in the corresponding PSTH bin).

Be sure to justify your answers to the following questions by showing any necessary calculations.

4. Would the average of the spike-triggered ensemble yield an unbiased estimate of $\mathbf{k}_1$ if the true joint response is indeed described by a 1D LN model? What can you say about the estimate of $\mathbf{k}_1$ if the true response is a 2D LN function? (6 marks)

5. Can the same weighted spike-triggered ensemble be used for STC analysis? (6 marks)

6. A simpler spike-triggered ensemble would have been obtained by just weighting $\mathbf{x}_t$ with the number of spikes generated in both cells. Explain why the two approaches might yield different subspaces. When would they be the same? (6 marks)

4 Learning

In class we saw that the structured part of the connectivity was weaker than the random part. For instance, in an all-inhibitory Hopfield network, the equations might look like

$$x_i(t+1) = \phi\left(\sqrt{n} I_0 - \frac{1}{\sqrt{n}} \sum_{j=1}^{n} w_{ij} x_j + \frac{\beta}{n} \eta_i^\mu \xi_j^\mu x_j(t)\right) \tag{1}$$

where w_{ij} is a random matrix with $\mathcal{O}(1)$ elements. The structured weights are, therefore, smaller than the random weights by a factor of $1/\sqrt{n}$. (That was the scaling in the networks question.) Since the structured weights are, presumably, the ones that are learned, this suggests that if n is large, during learning the change in weights will be small. We might, then, be able Taylor expand the weights around their initial values, which should simplify learning. We'll do that in a feedforward network, giving us the so-called neural tangent kernel. The goal is to show that when n is large, a first order Taylor expansion is a good approximation.

Consider a vanilla feedforward network,

$$\mathbf{x}^l = \phi(\mathbf{w}^l \cdot \mathbf{x}^{l-1}) \tag{2}$$ {ff}

where l labels layer and $\mathbf{w}^l$ is the *initial* weight. We'll use $f(\mathbf{w}, \mathbf{x})$ to refer to the whole network, so the mapping is

$$y = f(\mathbf{w}, \mathbf{x}) \tag{3}$$ {y=f}

where y is a scalar. All layers contain n neurons, and the last layer, layer L, is linear,

$$y = \mathbf{w}^{L+1} \cdot \mathbf{x}^L \,. \tag{4}$$ {output}

We want to minimize a loss function based on p training examples,

$$\{\mathbf{x}_\mu, y_\mu\},\ \mu = 1, ..., p \tag{5}$$

where $\mathbf{x}_\mu$ is the input at the bottom layer. We'll assume the weights change by $\Delta\mathbf{w}$, and choose $\Delta\mathbf{w}$ to minimize the square loss,

$$\mathcal{L} = \frac{1}{2} \sum_{\mu=1}^{p} \left(y_\mu - f(\mathbf{w} + \Delta\mathbf{w}, \mathbf{x}_\mu)\right)^2 . \tag{6}$$ {sqloss}

Note that $\Delta\mathbf{w}$ is shorthand for a change of weights in each layer.

$$\Delta\mathbf{w} = \left(\Delta\mathbf{w}^1, \Delta\mathbf{w}^2, ..., \Delta\mathbf{w}^{L+1}\right). \tag{7}$$

1. Show that if we Taylor expand to first order, learning becomes a linear regression problem,

$$\mathcal{L} \approx \frac{1}{2} \sum_{\mu=1}^{p} \left(\tilde{y}_\mu - \sum_l \tilde{\mathbf{x}}_\mu^l \cdot \Delta\mathbf{w}^l\right)^2 \tag{8}$$ {loss}

where {tildes}

$$\tilde{y}_\mu \equiv y_\mu - f(\mathbf{w}, \mathbf{x}_\mu) \tag{9a}$$

$$\tilde{\mathbf{x}}^l_\mu \equiv \frac{\partial f(\mathbf{w}, \mathbf{x}_\mu)}{\partial \mathbf{w}^l}\,. \tag{9b}$$

The dot product in Eq. (8) is shorthand, since $\tilde{\mathbf{x}}^l_\mu$ is really a matrix; it's defined to be

$$\tilde{\mathbf{x}}^l_\mu \cdot \Delta\mathbf{w}^l \equiv \text{trace}\left\{\tilde{\mathbf{x}}^l_\mu \cdot \Delta\mathbf{w}^{l\,T}\right\} \tag{10}$$

where here "$\cdot$" is now the usual dot product, and T denotes transpose.

(5 marks)

2. Assume that the total number of parameters (meaning the total number of weights) is large compared to p. In this case, we can fit the data perfectly. Show that if the weights are learned using gradient descent, then

$$\Delta\mathbf{w}^l = \frac{1}{n}\sum_\mu a^l_\mu \tilde{\mathbf{x}}^l_\mu \tag{11}$$ {w}

for some set of parameters, a^l_μ. The prefactor $1/n$ is for convenience only (recall that n is the size of each layer).

(5 marks)

3. Show that the a^l_μ satisfy the equation

$$\tilde{y}_\mu = \sum_{\nu,l} C^l_{\mu\nu} a^l_\nu \tag{12}$$ {ytilde}

where

$$C^l_{\mu\nu} = \frac{\tilde{\mathbf{x}}^l_\mu \cdot \tilde{\mathbf{x}}^l_\nu}{n}\,. \tag{13}$$ {Cdef}

(5 marks)

4. We now want to determine whether or not a first order Taylor expansion is valid. At the very least, this means $\Delta\mathbf{w}^l$ must be small compared to the initial weights, so the first thing we'll check is the size of $\Delta\mathbf{w}^l$. For that we'll need the typical size of the a^l_μ, which in turn requires that we need the typical size of $C^l_{\mu\nu}$. Since we expect $C^l_{\mu\nu}$ to be largest when $\mu = \nu$, we'll compute $C^l_{\mu\mu}$.

   Define the synaptic drive,

$$\mathbf{h}^l \equiv \mathbf{w}^l \cdot \mathbf{x}^{l-1}\,, \tag{14}$$ {hdef}

which means (via Eq. (2)) that

$$\mathbf{x}^l = \phi(\mathbf{h}^l)\,. \tag{15}$$ {xdef}

Show that

$$\tilde{\mathbf{x}}_\mu^l = \frac{\partial f(\mathbf{w}, \mathbf{x}_\mu)}{\partial \mathbf{h}^l}\, \mathbf{x}_\mu^{l-1} \tag{16}$$ {xll}

where, as usual, two adjacent vectors refers to an outer product. Consequently,

$$C_{\mu\mu}^l = \left| \frac{\partial f(\mathbf{w}, \mathbf{x}_\mu)}{\partial \mathbf{h}^l} \right|^2 \frac{|\mathbf{x}_\mu^{l-1}|^2}{n}\,. \tag{17}$$ {C}

(5 marks)

5. We'll assume that the activity in each layer is $\mathcal{O}(1)$, so the second term in the above expression is $\mathcal{O}(1)$. We just need to figure out how big the first term is. To determine that, show that we can backprop that term,

$$\frac{\partial f(\mathbf{w}, \mathbf{x}_\mu)}{\partial \mathbf{h}^l} = \frac{\partial f(\mathbf{w}, \mathbf{x}_\mu)}{\partial \mathbf{h}^{l+1}} \cdot \mathbf{w}^{l+1} \odot \phi'(\mathbf{h}^l) \tag{18}$$ {backprop}

where "$\odot$" indicates element-wise multiplication,

$$(\mathbf{w}^{l+1} \odot \phi'(\mathbf{h}^l))_{ij} \equiv w_{ij}^{l+1} \phi'(h_j^l)\,. \tag{19}$$

(5 marks)

6. To proceed, we need to know how big the weights are initially. We'll assume that they're random, and scale as $1/\sqrt{n}$,

$$w_{ij}^l \sim \frac{1}{\sqrt{n}}\,. \tag{20}$$

Why is this scaling reasonable? To answer this, you might consider what would happen if the weights were $\mathcal{O}(1/n)$ or $\mathcal{O}(1)$. But there are other approaches.

(5 marks)

7. Because the initial weights are random we expect the terms in the dot product on the right hand side of Eq. (18) to be independent of each other. Using this fact, show that

$$\frac{\partial f(\mathbf{w}, \mathbf{x}_\mu)}{\partial \mathbf{h}^{l+1}} \cdot \mathbf{w}^{l+1} \odot \phi'(\mathbf{h}^l) \sim \left(\text{typical size of } \frac{\partial f(\mathbf{w}, \mathbf{x}_\mu)}{\partial h_i^{l+1}} \right)\,. \tag{21}$$

Consequently,

$$\frac{\partial f(\mathbf{w}, \mathbf{x}_\mu)}{\partial h_i^l} \sim \frac{\partial f(\mathbf{w}, \mathbf{x}_\mu)}{\partial h_i^L}\,. \tag{22}$$

to finish the calculation, show that

$$\frac{\partial f(\mathbf{w}, \mathbf{x}_\mu)}{h_i^L} \sim \frac{1}{\sqrt{n}} \tag{23}$$

Consequently, $\partial f(\mathbf{w}, \mathbf{x}_\mu)/\partial h_i^l \sim 1/\sqrt{n}$ in all layers. This in turn implies that $C_{\mu\mu} \sim \mathcal{O}(1)$ (see Eq. (17), which implies that the a_μ^l are $\mathcal{O}(1)$.

To determine the size of $\Delta \mathbf{w}^l$, note that the elements of $\tilde{\mathbf{x}}_\mu^l$ are about the same size as the elements of $\partial f(\mathbf{w}, \mathbf{x}_\mu)/\partial \mathbf{h}^l$, which follows from Eq. (16) and the fact that $x_i^l \sim \mathcal{O}(1)$. Consequently, via Eq. (11),

$$\Delta w_{ij}^l \sim \mathcal{O}\left(\frac{1}{n^{3/2}}\right) . \tag{24}$$

Since the initial weights are $\mathcal{O}(1/\sqrt{n})$, it follows that $|\Delta w_{ij}^l| \ll |w_{ij}|$. As an aside, it turns out that $\Delta w_i^{L+1} \sim \mathcal{O}(1/n)$, so changes to those weights are larger than the changes to the weights in all the intermediate layers.

(10 marks)

This does not mean the Taylor expansion is valid, but that's actually not hard to show. Only about 5 pages of algebra. ;)

Solutions

1. Taylor expanding Eq. (3) gives

$$y = f(\mathbf{w} + \Delta\mathbf{w}, \mathbf{x}) \approx f(\mathbf{w}, \mathbf{x}) + \sum_l \frac{\partial f(\mathbf{w}, \mathbf{x})}{\partial \mathbf{w}^l} \cdot \Delta \mathbf{w}^l . \tag{25}$$

Thus, the loss, Eq. (6), is

$$\mathcal{L} \approx \frac{1}{2} \sum_{\mu=1}^{p} \left(y_\mu - f(\mathbf{w}, \mathbf{x}_\mu) - \sum_l \frac{\partial f(\mathbf{w}, \mathbf{x}_\mu)}{\partial \mathbf{w}^l} \cdot \Delta \mathbf{w}^l . \right)^2 \tag{26}$$

Using the definitions in Eq. (9), we recover Eq. (8).

2. First, the naive answer which is what I had in mind when I wrote the question: Under gradient descent on the approximate loss function, the weights evolve according to

$$\Delta(\Delta \mathbf{w}^l) = -\eta \frac{\partial \mathcal{L}}{\partial \mathbf{w}^l} = \eta \sum_{\mu=1}^{p} \left(\tilde{y}_\mu - \sum_l \tilde{\mathbf{x}}_\mu^l \cdot \Delta \mathbf{w}^l \right) \tilde{\mathbf{x}}_\mu^l . \tag{27}$$

Consequently, weight changes in layer l are constrained to lie in the subspace spanned by the $\tilde{\mathbf{x}}_\mu^l$. This gives us Eq. (11).

But, as pointed out to me by Francesca Mastrogiuseppe, this is in fact not correct. Let's start instead with stochastic gradient descent (although what we'll say applies to gradient descent as well), with different learning rates in each layer. We'll use t to label update, so we write

$$\Delta(\Delta \mathbf{w}_t^l) = \eta^l \epsilon_{\mu_t}\big(\Delta \mathbf{w}(t-1)\big) \tilde{\mathbf{x}}_{\mu_t}^l \tag{28}$$

where

$$\epsilon_\mu(\Delta \mathbf{w}) \equiv \tilde{y}_\mu - \sum_l \tilde{\mathbf{x}}_\mu^l \cdot \Delta \mathbf{w}^l . \tag{29}$$

After T updates, $\Delta\mathbf{w}$ is given by

$$\Delta\mathbf{w}^l \equiv \Delta\mathbf{w}^l_T = \eta^l \sum_{t=1}^{T} \epsilon_{\mu_t}\big(\Delta\mathbf{w}(t-1)\big)\tilde{\mathbf{x}}^l_{\mu_t} = \eta^l \sum_{\mu} \left(\sum_{t=1}^{T} \epsilon_{\mu_t}\big(\Delta\mathbf{w}(t-1)\big)\delta_{\mu,\mu_t} \right) \tilde{\mathbf{x}}^l_{\mu} \tag{30}$$ {a_mu_correc

where $\delta_{\mu\mu_t}$ is the usual Kronecker delta. Now make the definition

$$a_\mu \equiv n \left(\sum_l \eta^l \right) \left(\sum_{t=1}^{T} \epsilon_{\mu_t}\big(\Delta\mathbf{w}(t-1)\big)\delta_{\mu,\mu_t} \right) \tag{31}$$

so that

$$\Delta\mathbf{w}^l = \frac{1}{n} \sum_\mu a_\mu \gamma^l \tilde{\mathbf{x}}^l_\mu \tag{32}$$ {dwfinal}

where γ^l is the relative learning rate,

$$\gamma^l \equiv \frac{\eta^l}{\sum_m \eta^m} \,. \tag{33}$$

The point here is that the a_μ don't depend on layer.

3. First, the naive answer: Because we can fit the training examples perfectly, we have

$$\tilde{y}_\mu = \sum_l \tilde{\mathbf{x}}^l_\mu \cdot \Delta\mathbf{w}^l = \sum_l \tilde{\mathbf{x}}^l_\mu \cdot \frac{1}{n} \sum_\nu a^l_\nu \tilde{\mathbf{x}}^l_\nu = \sum_\nu \frac{\tilde{\mathbf{x}}^l_\mu \cdot \tilde{\mathbf{x}}^l_\nu}{n} a^l_\nu \,. \tag{34}$$

Using Eq. (13) for $C^l_{\mu\nu}$, we recover Eq. (12).

Now the correct answer, where we use Eq. (**??**) rather than Eq. (11). Again we fit the training examples perfectly (for which we need $T \to \infty$ in Eq. (**??**)), so we have

$$\tilde{y}_\mu = \sum_l \tilde{\mathbf{x}}^l_\mu \cdot \frac{1}{n} \sum_\nu a_\nu \gamma^l \tilde{\mathbf{x}}^l_\nu = \sum_l \gamma^l \sum_\nu \frac{\tilde{\mathbf{x}}^l_\mu \cdot \tilde{\mathbf{x}}^l_\nu}{n} \, a_\nu \,. \tag{35}$$

Defining $C_{\mu\nu}$ (without a superscript l this time!),

$$C_{\mu\nu} \equiv \sum_l \gamma^l \frac{\tilde{\mathbf{x}}^l_\mu \cdot \tilde{\mathbf{x}}^l_\nu}{n} \,, \tag{36}$$

the equation for a_μ is

$$\tilde{y}_\mu = \sum_\nu C_{\mu\nu} a_\nu \,. \tag{37}$$

This changes some of the details of the rest of the calculation, but still the main thing we want to compute is $\mathbf{x}^l_\mu \cdot \mathbf{x}^l_\mu$.

4. By definition,

$$\tilde{\mathbf{x}}_\mu^l = \frac{\partial f(\mathbf{w}, \mathbf{x}_\mu)}{\partial \mathbf{w}^l} \,. \tag{38}$$

By the chain rule,

$$\frac{\partial f(\mathbf{w}, \mathbf{x}_\mu)}{\partial \mathbf{w}^l} = \frac{\partial f(\mathbf{w}, \mathbf{x}_\mu)}{\partial \mathbf{h}^l} \cdot \frac{\partial \mathbf{h}^l}{\partial \mathbf{w}^l} \,. \tag{39}$$ {dfdw}

Using Eq. (14), we see that

$$\frac{\partial \mathbf{h}^l}{\partial \mathbf{w}^l} = \mathbf{I}\mathbf{x}^{l-1} \tag{40}$$

where $\mathbf{I}$ is the identity matrix. (To really convince yourself of this, it helps to use indices, and note that $\partial \mathbf{h}^l / \partial \mathbf{w}^l$ has three of them!) Inserting this into Eq. (**??**), we recover Eq. (16).

5. As usual, we have

$$\frac{\partial f(\mathbf{w}, \mathbf{x}_\mu)}{\partial \mathbf{h}^l} = \frac{\partial f(\mathbf{w}, \mathbf{x}_\mu)}{\partial \mathbf{h}^{l+1}} \cdot \frac{\partial \mathbf{h}^{l+1}}{\partial \mathbf{h}^l} \cdot \,. \tag{41}$$

To compute the last term, we combine Eq. (14) with (15); that yields Eq. (18).

6. The term that goes inside the nonlinearity in layer l is

$$(\mathbf{w}^l \mathbf{x}^{l-1})_i = \sum_{j=1}^{n} w_{ij}^l x_j^{l-1} \,. \tag{42}$$

The weights are chosen randomly (and zero mean), so the sum on the right hand side scales as $\sqrt{n}\times$ (typical size of w_{ij}^l) $\times$ (typical size of x_j^{l-1}). Because the activity, x_j^{l-1}, must be $\mathcal{O}(1)$, if we want the sum to be $\mathcal{O}(1)$ then w_{ij}^l must scale as $1/\sqrt{n}$.

7. We have

$$\left(\frac{\partial f(\mathbf{w}, \mathbf{x}_\mu)}{\partial \mathbf{h}^{l+1}} \cdot \mathbf{w}^{l+1} \odot \phi'(\mathbf{h}^l) \right)_i = \sum_{j=1}^{n} \frac{\partial f(\mathbf{w}, \mathbf{x}_\mu)}{\partial h_i^{l+1}} \, w_{ij}^{l+1} \odot \phi'(h_j^l) \,. \tag{43}$$

Using the same reasoning as above, the sum scales as the typical size of $\partial f(\mathbf{w}, \mathbf{x}_\mu)/\partial h_i^{l+1}$.

In the output layer,

$$y_\mu = f(\mathbf{w}, \mathbf{x}_\mu) = \mathbf{w}^{L+1} \phi(\mathbf{h}^L) \tag{44}$$

(see Eq. (4)). Consequently,

$$\frac{\partial f(\mathbf{w}, \mathbf{x}_\mu)}{\partial h_i^L} = w_i^{L+1} \phi'(h_i^L) \sim \frac{1}{\sqrt{n}} \,. \tag{45}$$