Gatsby Computational Neuroscience Unit
Theoretical Neuroscience

Final examination, theoretical neuroscience
21 May 2025

Part I – short questions

There are four sections with four questions each. Please answer three out of each four (**noting special instructions for Networks**), starting the answers for each section on a new page. Don't forget to write your name at the top of each block of answers.

Good luck!

1 Biophysics

1. **Synaptic depression.** Consider a synapse that exhibits synaptic depression, meaning its release probability, p, evolves according to

$$\frac{dp}{dt} = \frac{p_0 - p}{\tau} - \sum_n \delta(t - t_n) p(t_n) \xi_n \tag{1}$$

where t_n is the time of the presynaptic spike and ξ_n is a random variable that is either 0 or 1,

$$\xi_n = \begin{cases} 1 & \text{probability } p(t_n) \\ 0 & \text{probability } 1 - p(t_n)\,. \end{cases} \tag{2}$$

A neuron is firing at frequency ν (let's say it's Poisson, although that doesn't really matter). In the limit $\nu\tau \gg 1$, what's the equilibrium release probability?

Solution When $\nu\tau \gg 1$, the release probability is updated a large number of times relative to how fast it decays. In that regime, ξ_n can be replaced by it's average, p, giving us the equation

$$\frac{dp}{dt} \approx \frac{p_0 - p}{\tau} - \nu p^2\,. \tag{3}$$

In steady state, p is given by the solution to

$$\nu\tau p^2 + p - p_0 = 0\,. \tag{4}$$

The positive solution, which is the relevant one, is

$$p = \frac{\sqrt{1 + 4p_0\nu\tau} - 1}{2\nu\tau} \approx \sqrt{\frac{p_0}{\nu\tau}}\,. \tag{5}$$

2. **H-current.** The H-current is a sodium current that activates at low voltages (around -60 mV) and de-activates slowly at high voltages (around -50). Explain why this current can lead to bursting.

Solution If a neuron is not firing, it sits at around -65 mV – low enough for the H-current to activate. Because the H-current is a sodium current, it's inward, and so raises the membrane potential. When the membrane potential goes up enough, the neuron starts firing repetitively. The repetitive firing raises the voltage, causing the H-current to inactivate, so eventually the neuron stops firing. These alternating periods of firing and silence corresponds to bursting.

3. **NMDA channels.** Explain how NMDA channels act as coincidence detectors of pre and postsynaptic spikes. Why is this important for synaptic plasticity?

Solution The NMDA channel is mediated by sodium, so its reversal potential is around 0 mV. If it were a standard channel, it would conduct current whenever there was a pre-synaptic spike. However, at voltages less than around -65 mV, it's blocked by magnesium. So it's open only at high voltages, when the magnesium block is removed. At a synapse, voltage is raised mainly by a back-propagating action potential. Consequently, there's an NMDA-mediated current only when there's a coincidence: both a pre and postsynaptic spike.

This is important because we believe that effective learning requires a coincidence between pre and postsynaptic firing.

4. **Refractory period.** The current through an active potassium channel is proportional $n^4(V - \mathcal{E}_K)$. Write down the dynamics of the n-channel, and the value of $\mathcal{E}_k$, that will lead to a refractory period.

Solution A neuron will exhibit a refractory period if the potassium reversal potential is negative (for instance, $\mathcal{E}_k = -80$ mV) and the potassium channel opens at the high voltages produced by a spike. The latter is ensured if n obeys an equation like

$$\tau \frac{dn}{dt} = \sigma\left(\frac{V - V_0}{\Delta V}\right) - n \tag{6}$$

where σ is the sigmoid function, $V_0 \sim -50$ mV, $\Delta V \sim 15mV$, and $\tau \sim 1 - 2$ ms.

2 Networks

In this section, you should attempt questions 2 and 3. You may skip either 1 or 4. (In other words, attempt 30 marks worth.)

1. (10 marks) Consider a network of one excitatory and one inhibitory populations which have a stable fixed point given by

$$\nu_E = \phi_E\big(W_{EE}\,\nu_E - W_{EI}\nu_I + h_E\big) \tag{1a}$$
$$\nu_I = \phi_I\big(W_{IE}\,\nu_E - W_{II}\nu_I + h_I\big) \tag{1b}$$

where ϕ_E and ϕ_E are monotonically increasing functions, and all weights are positive. Under which conditions the inhibitory firing rate decreases with a positive perturbation to its inputs (i.e. $h_I \to h_I + \delta h_I$ where δh_I is infinitesimally small)? What happens to excitatory firing rate?

Solution

The need to compute the linear response. We define $\Phi'_{\alpha\alpha} = \phi'_\alpha\big(W_{\alpha E}\,\nu^*_E - W_{\alpha I}\nu^*_I + h_\alpha\big)$

$$\frac{d\nu_\alpha}{dh_\beta} = \Phi'_{\alpha\alpha}\big(W_{\alpha E}\,\frac{d\nu_E}{dh_\beta} - W_{I\alpha}\frac{d\nu_I}{dh_\beta} + \delta_{\alpha\beta}\big) \tag{2a}$$

Defining $M_{\alpha\beta} = \frac{d\nu_E}{dh_\beta}$, the above becomes

$$M = \Phi'\big(WM + I\big) \tag{3a}$$
$$M = (I - \Phi' W)^{-1}\Phi' \tag{3b}$$
$$M = (-J)^{-1}\Phi' \tag{3c}$$

Where J is the Jacobian of the system.

$$M = \begin{pmatrix} 1 - \Phi'_{EE}W_{EE} & \Phi'_{EE}W_{EI} \\ -\Phi'_{IE}W_{EI} & 1 + \Phi'_{II}W_{II} \end{pmatrix}^{-1} \begin{pmatrix} \Phi'_{EE} & 0 \\ 0 & \Phi'_{II} \end{pmatrix} \tag{4}$$

$$M = \frac{1}{|-J|} \begin{pmatrix} 1 + \Phi'_{II}W_{II} & -\Phi'_{EE}W_{EI} \\ \Phi'_{IE}W_{EI} & 1 - \Phi'_{EE}W_{EE} \end{pmatrix} \begin{pmatrix} \Phi'_{EE} & 0 \\ 0 & \Phi'_{II} \end{pmatrix} \tag{5}$$

Because $|-J| > 0$ for linear stability,

$$\frac{d\nu_I}{dh_I} \propto 1 - \Phi'_{EE}W_{EE} \tag{6a}$$
$$\frac{d\nu_E}{dh_I} \propto -\Phi'_{EE}W_{EI} \tag{6b}$$
$$\tag{6c}$$

And the excitatory activity always decreases.

2. (5 marks) Consider a large-scale network of excitatory and inhibitory neurons, in which the firing rate of one excitatory, ν^E_i (or inhibitory, ν^E_i) neuron i is given by:

$$\nu^E_i = \phi_E\big(\sum_j J^{EE}_{ij}\,\nu^E_j - J^{EI}_{ij}\nu^I_j + h^E\big) \tag{7a}$$
$$\nu^I_i = \phi_I\big(\sum_j J^{IE}_{ij}\,\nu^E_j - J^{II}_{ij}\nu^I_j + h^I\big) \tag{7b}$$

and in which $J_{ij}^{\alpha\beta} = \frac{J^{\alpha\beta}}{\sqrt{K}}$ with probability $p = \frac{K}{N}$ where N is the size of each population and $h^\alpha = \sqrt{K} h_0^\alpha$ with h_0^E a size independent constant, for $\alpha, \beta \in \{E, I\}$.

Show that in the large K limit, the mean rates track inputs linearly.

Solution If they overcomplicate themselves they will re-derive the mean-field equations I derived in class

$$z^\alpha = \sum_\beta J^{\alpha\beta}\sqrt{K^\beta}\nu^\beta + h^\alpha \tag{8a}$$

$$\Delta_z^\alpha = \sum_\beta (J^{\alpha\beta})^2 (v_\nu^\beta - p^\beta {\nu^\beta}^2) \tag{8b}$$

$$\nu^\alpha = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \phi(z^\alpha + \sqrt{\Delta_z^\alpha} z) e^{-z^2/2} \tag{8c}$$

$$v_\nu^\alpha = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \phi(z^\alpha + \sqrt{\Delta_z^\alpha} z)^2 e^{-z^2/2} \tag{8d}$$

If not, they will just compute Eq. (11a). In any case,

$$\sum_\beta J^{\alpha\beta}\nu^\beta + h_0^\alpha = \frac{z^\alpha}{\sqrt{K}} \to 0 \tag{9}$$

and hence the mean firing rates ν^β follow the inputs linearly.

3. (15 marks) Consider the low-rank network

$$\dot{\mathbf{x}} = -\mathbf{x} + \mathbf{W}\phi(\mathbf{x}) \tag{10}$$

where $\mathbf{x} \in \mathbb{R}^N$, $\mathbf{W} \in \mathbb{R}^{N\times N}$ and $\mathbf{W} = \frac{\mathbf{m}\mathbf{n}^T}{N}$, where $\mathbf{m}$, $\mathbf{n} \in \mathbb{R}^N$.

   (a) What is the direction of the fixed point in this network?
   Solution $\mathbf{x} = \text{scalar} * \mathbf{m}$, so the direction is $\mathbf{m}$

   (b) We define $\kappa = \frac{\mathbf{n}^T\phi(\mathbf{x})}{N}$. If we assume that the elements of $\mathbf{m}$ and $\mathbf{n}$, which we now call m and n, are independently sampled from a joint Gaussian distribution. In the limit of large N, why can we write $\kappa = \mathbb{E}_{m,n\sim P(m,n)}[n\phi(\kappa m)]$?
   Solution Law of large numbers

   (c) We now take $m = \sigma_m y$ with $y \sim \mathcal{N}(0,1)$ and σ_m a the standard deviation of m. We also take $n = \sigma_n\left(\rho y + \sqrt{1-\rho^2} z\right)$ with $z \sim \mathcal{N}(0,1)$, with ρ the correlation between the variables m and n and σ_n the standard deviation of n. Compute a self-consistent solution for κ using the hint below.
   Hint: $\int_{-\infty}^{\infty} \frac{d}{dz}\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} f(\alpha z)\right) dz$=0. Do parts integration to obtain (a simpler/modified version of) Stein's Lemma
   Solution If we define $\mathcal{D}z = \exp(-\frac{z^2}{2})dz$, we can write

$$\kappa = \int\int \mathcal{D}z\, \mathcal{D}y\, n\phi(km) \tag{11a}$$

$$\kappa = \int\int \mathcal{D}z\, \mathcal{D}y\, \sigma_n\left(\rho y + \sqrt{1-\rho^2} z\right)\phi(\kappa\sigma_m y) \tag{11b}$$

$$\kappa = \sigma_n\rho \int \mathcal{D}y\, y\phi(\kappa\sigma_m y) + \sigma_n\left(\int \mathcal{D}z\, z\right)\int \mathcal{D}y\, \sqrt{1-\rho^2}\phi(\kappa\sigma_m y) \tag{11c}$$

$$\kappa = \sigma_n\rho \int \mathcal{D}y\, y\phi(\kappa\sigma_m y) \qquad \text{Use steins lemma for the next step} \tag{11d}$$

$$\kappa = \sigma_n\kappa\sigma_m\rho \int \mathcal{D}y\, \phi'(\kappa\sigma_m y) \tag{11e}$$

$$\frac{1}{\rho} = \sigma_n\sigma_m \langle\phi'(\kappa\sigma_m y)\rangle_{y\sim\mathcal{N}(0,1)} \tag{11f}$$

$$\tag{11g}$$

(d) Assume that $\langle \phi'(\kappa y) \rangle_{y \sim \mathcal{N}(0,1)}$ is bell shaped as a function of κ. If we take $\sigma_n = \sigma_m = 1$, how many values of κ are solutions as we increase ρ?

Solution For small ρ we have no solution, and for large enough ρ we have 2.

4. (10 marks) Consider an almost standard Hopfield network,

$$S_i(t+1) = \text{sign}\left[\frac{1}{n}\sum_{j=1}^{n} J_{ij} S_j(t)\right] . \tag{12}$$

The connectivity matrix is given, as usual, by

$$J_{ij} = \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu \tag{13}$$

where

$$\xi_i^\mu = \begin{cases} +1 & \text{probability } 1/2 \\ -1 & \text{probability } 1/2 . \end{cases} \tag{14}$$

Now the almost standard part: the ξ_i^μ are correlated; for $\mu \neq \nu$,

$$\text{probability}(\xi_i^\mu = \xi_i^\nu) = q \tag{15}$$

with $q > 1/2$.

Show that in the large n limit, it's possible to embed memories (fixed points of the update rule, Eq. (12)) so long as

$$p \ll 1 + \frac{1}{(2q-1)^2} . \tag{16}$$

Solution As usual, we'll let $S_i = \xi_i^\mu$, and see if that's a fixed point. This yields

$$\xi_i^\mu \stackrel{?}{=} \text{sign}\left[\xi_i^\mu + \sum_{\nu \neq \mu} \xi_i^\nu \frac{1}{n} \sum_j \xi_j^\nu \xi_j^\mu\right]$$

In the large n limit,

$$\frac{1}{n}\sum_j \xi_j^\nu \xi_j^\mu = q \times (+1) + (1-q) \times (-1) = 2q - 1 ,$$

giving us

$$\xi_i^\mu \stackrel{?}{=} \text{sign}\left[\xi_i^\mu + (2q-1)\sum_{\nu \neq \mu} \xi_i^\nu\right]$$

The variance of the sum over ν is $p-1$, and so we may write

$$\xi_i^\mu \stackrel{?}{=} \text{sign}\left[\xi_i^\mu + (2q-1)(p-1)^{1/2}\zeta_i\right]$$

where ζ_i is a zero mean, unit variance random variable. Assuming ζ_i isn't heavy tailed (it isn't), we have equality if

$$(2q-1)(p-1)^{1/2} \ll 1$$

which we may write

$$p \ll 1 + \frac{1}{(2q-1)^2} .$$

3 Coding

1. A simple "coincidence detector" neuron fires whenever it receives two input spikes within time interval τ. Assume that the output spike is coincident with the second of the two inputs. After it fires, the neuron resets its state instantaneously, losing any memory of the preceding spikes. Suppose the neuron receives input through two synapses, and that each presynaptic spike train is homogenous Poisson with rate λ.

   (a) What is the firing rate of the coincidence detector? Check that your answer makes sense for very small and very large τ.

   (b) What is the ISI distribution of the output spike train? [You may leave the answer in integral form.]

   (c) How do your answers change if the neuron only fires when the two coincident inputs come from different synapses?

   Solution

   (a) Call the output rate ρ, and consider counting spikes within a very long time window T. For each input spike, the probability that a second one arrives within time τ is $1 - e^{-2\lambda\tau}$. There are ρT such pairs on average (by definition of ρ). Since the second spikes themselves can't start a pair, there are $2\lambda T - \rho T$ inputs (on average) that can. Consistency requires

   $$\rho T = (2\lambda T - \rho T)(1 - e^{-2\lambda\tau})$$
   $$\Rightarrow \rho = 2\lambda \frac{1 - e^{-2\lambda\tau}}{2 - e^{-2\lambda\tau}}$$

   For small τ this approaches 0. For large τ it approaches λ, which is correct as every other input will generate an output.

   (b) Following an output, the first of the next coincident input pair is obtained by considering a thinned Poisson process. The time to this is exponential with time constant $1/\rho$. The additional time (δ) to the second in the pair (and thus the next output) is a truncated exponential with time constant $1/2\lambda$. So the ISI density is given by

   $$p(\Delta) = \int_0^{\min(\Delta,\tau)} d\delta\, \rho e^{-\rho(\Delta-\delta)} \frac{2\lambda e^{-2\lambda\delta}}{1 - e^{-2\lambda\tau}}$$

   Note the upper limit that ensures that both the intervals in the sum remain positive.

   (c) The rate of candidate first spikes is the same. But it will be the first of a coincident pair if both (a) a spike arrives on the other synapse within τ and (b) no spikes arrive on the same synapse first (if they do, the pair will started by that later spike). The probability of both happening is

   $$\pi = \int_0^{\tau} dt\, \lambda e^{-\lambda t}(e^{-\lambda t}) = \frac{1}{2}(1 - e^{-2\lambda\tau})$$

   (Indeed, another way to look at it is that the pairs computed in part (a) are equally likely to comprise same-synapse and cross-synapse events. So the probability is half the previous value.) So

   $$\rho = 2\lambda \frac{1 - e^{-2\lambda\tau}}{3 - e^{-2\lambda\tau}}$$

   For the ISI, the thinning logic applies with the new ρ, and the added time for the second spike now depends on an exponential process with time constant $1/\lambda$, so

   $$p(\Delta) = \int_0^{\min(\Delta,\tau)} d\delta\, \rho e^{-\rho(\Delta-\delta)} \frac{\lambda e^{-\lambda\delta}}{1 - e^{-\lambda\tau}}$$

2. We model the firing rate responses of two neurons to an image stimulus s as normal:

   $$\begin{pmatrix} r_1 \\ r_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1(s) \\ \mu_2(s) \end{pmatrix}, \begin{pmatrix} \sigma_1 & \rho \\ \rho & \sigma_2 \end{pmatrix} \right)$$

   We observe that, as we draw s randomly from a collection of images, $\mu_1(s)$ and $\mu_2(s)$ appear to be marginally independent and vary over a much larger range than $\sqrt{\sigma_1}$ and $\sqrt{\sigma_2}$.

(a) Suggest a conclusion that might be drawn about the coding properties of the neurons based on the observed marginal independence.

(b) Find a condition involving (possibly just some of) the distributions of μ_1 and μ_2 and the (constant) noise terms σ_1, σ_2 and ρ for the coding to be synergistic.

Solution

(a) The neurons code sufficiently different (and thus independent) parts of the image – most likely they have receptive fields that are well separated in space and/or some feature dimension.

(b) The condition is just $\rho \neq 0$. We are told that μ_i vary over large ranges, so can assume that the total entropies are dominated by the (independent) μ variance. Thus $\mathbf{H}[r_1, r_2] \approx \mathbf{H}[r_1] + \mathbf{H}[r_2]$. So synergy requires that the joint noise entropy is smaller than the sum of marginals. We have

$$\begin{aligned}
\frac{1}{2}\log\left|2\pi e\begin{pmatrix}\sigma_1 & \rho\\ \rho & \sigma_2\end{pmatrix}\right| &= \frac{1}{2}\log\left((2\pi e)^2(\sigma_1\sigma_2 - \rho^2)\right)\\
&= \frac{1}{2}\log\left((2\pi e)^2(\sigma_1\sigma_2)(1 - \frac{\rho^2}{\sigma_1\sigma_2})\right)\\
&= \frac{1}{2}\log 2\pi e\sigma_1 + \frac{1}{2}\log 2\pi e\sigma_2 + \frac{1}{2}\log\left(1 - \frac{\rho^2}{\sigma_1\sigma_2}\right)\\
&< \frac{1}{2}\log 2\pi e\sigma_1 + \frac{1}{2}\log 2\pi e\sigma_2
\end{aligned}$$

whenever $\rho \neq 0$

3. Recall that the maximally informative dimensions (MID) approach to characterising neural responses uses a histogram-based method to look for stimulus directions along which the projected spike-triggered stimulus ensemble (STE) differs most (by KL) from the total stimulus ensemble (TSE). An alternative to the histogram approach is to use a simple model of the STE – for example, a multivariate normal, which leads to an approach called iSTAC.

(a) Let the TSE be $\mathbf{s} \sim \mathcal{N}(0, I)$ and the maximum-likelihood normal fit to the STE be $\mathbf{s}|\text{spike} \sim \mathcal{N}(\mu, \Lambda)$. We want to find a subspace described by a rectangular matrix K (with $K^\mathsf{T}K = I$) such that $\mathbf{KL}\big[\text{STE}(K^\mathsf{T}\mathbf{s})\big\|\text{TSE}(K^\mathsf{T}\mathbf{s})\big]$ is maximised. Formulate this objective in terms of μ, Λ and K.

(b) Show that if $\Lambda = \lambda I$, the optimal 1D subspace is aligned with the STA.

(c) Show that if $\mu = 0$, the optimal subspace is aligned with the STC-defined subspace—that is, it spans those dimensions where the STC is significantly different to the total stimulus covariance.

Solution

(a) The projected distributions are:

$$\begin{aligned}
\text{TSE:}&\quad \mathcal{N}(0, I)\\
\text{STE:}&\quad \mathcal{N}\big(K^\mathsf{T}\mu, K^\mathsf{T}\Lambda K\big)
\end{aligned}$$

So the KL divergence is

$$\begin{aligned}
\mathbf{KL}[\text{STE}\|\text{TSE}] &= \left\langle \frac{1}{2}\log|2\pi I| + \frac{1}{2}\mathbf{s}_K^\mathsf{T}\mathbf{s}_K\right\rangle_{\text{STE}} - \left\langle \frac{1}{2}\log\left|2\pi K^\mathsf{T}\Lambda K\right| + \frac{1}{2}(\mathbf{s}_K - K^\mathsf{T}\mu)^\mathsf{T}(K^\mathsf{T}\Lambda K)^{-1}(\mathbf{s}_K - K^\mathsf{T}\mu)\right\rangle_{\text{STE}}\\
&= \frac{1}{2}\log|2\pi I| + \frac{1}{2}\mathsf{Tr}\big[\langle \mathbf{s}_K\mathbf{s}_K^\mathsf{T}\rangle\big] - \frac{1}{2}\log\left|2\pi K^\mathsf{T}\Lambda K\right| - \frac{1}{2}\mathsf{Tr}\big[(K^\mathsf{T}\Lambda K)^{-1}\langle(\mathbf{s}_K - K^\mathsf{T}\mu)(\mathbf{s}_K - K^\mathsf{T}\mu)^\mathsf{T}\rangle\big]\\
&= \frac{1}{2}\mathsf{Tr}\big[K^\mathsf{T}\Lambda K + K^\mathsf{T}\mu\mu^\mathsf{T}K\big] - \frac{1}{2}\log\left|K^\mathsf{T}\Lambda K\right| - \frac{1}{2}\mathsf{Tr}\big[(K^\mathsf{T}\Lambda K)^{-1}(K^\mathsf{T}\Lambda K)\big]\\
&= \frac{1}{2}\mathsf{Tr}\big[K^\mathsf{T}(\Lambda + \mu\mu^\mathsf{T})K - I\big] - \frac{1}{2}\log\left|K^\mathsf{T}\Lambda K\right|
\end{aligned}$$

(b) Recalling that $K^\mathsf{T}K = I$, the objective becomes

$$\mathbf{KL}[\text{STE}\|\text{TSE}] = \frac{1}{2}\mathsf{Tr}\big[K^\mathsf{T}\mu\mu^\mathsf{T}K + (\lambda - 1)I\big] - \frac{1}{2}\log|\lambda I|$$

which is maximised when K aligns with μ.

(c) Now the objective is

$$\mathbf{KL}[\text{STE}\|\text{TSE}] = \frac{1}{2}\mathsf{Tr}\left[K^\mathsf{T}\Lambda K\right] - \frac{1}{2}\log\left|K^\mathsf{T}\Lambda K\right| - k$$

(k is the subspace dimensionality). If K selects an eigenvector of Λ with eigenvalue λ_i, the corresponding contribution to the objective is $\lambda_i - \log\lambda_i - 1$. This will be zero for $\lambda_i = 1$ (corresponding to dimensions in which the STC is the same as the TSE covariance). That value is the minimum of the function, so values of λ_i either greater than or smaller than 1 will contribute to the optimum.

4. Consider a linear noiseless model of an input-driven population

$$\dot{\mathbf{r}}(t) = A\mathbf{r}(t) + B\mathbf{u}(t)\,.$$

Suppose that $\mathbf{u}(t) \overset{\text{iid}}{\sim} \mathcal{N}(0, I)$. Assuming that the undriven model is stable, we want to find the long-run covariance of the activity.

(a) Let $\mathbf{r}(0) = 0$. Show that the covariance $W(t)$ of activity at time t is given by

$$W(t) = \int_0^t d\tau\, e^{A\tau}BB^\mathsf{T}e^{A^\mathsf{T}\tau}$$

(b) Show that

$$\int_0^t d\tau \left(\frac{d}{d\tau}e^{A\tau}BB^\mathsf{T}e^{A^\mathsf{T}\tau}\right) = AW(t) + W(t)A^\mathsf{T}$$

(c) Hence show that $W(\infty)$ satisfies $AW(\infty) + W(\infty)A^\mathsf{T} = -BB^\mathsf{T}$.

(d) $W(\infty)$ is called the (infinite-time) *controllability Gramian*. Why? What can we say if $W(\infty)$ has any zero eigenvalues?

Solution

(a) The state at time t is $\int_0^t d\tau e^{A\tau}B\mathbf{u}(t-\tau)$. So the covariance is

$$\begin{aligned}
W(t) &= \left\langle \left(\int_0^t d\tau e^{A\tau}B\mathbf{u}(t-\tau)\right)\left(\int_0^t d\tau e^{A\tau}B\mathbf{u}(t-\tau)\right)^\mathsf{T}\right\rangle \\
&= \int_0^t\int_0^t d\tau\, d\tau'\, e^{A\tau}B\left\langle\mathbf{u}(t-\tau)\mathbf{u}(t-\tau')^\mathsf{T}\right\rangle B^\mathsf{T}e^{A^\mathsf{T}\tau'} \\
&= \int_0^t\int_0^t d\tau\, d\tau'\, e^{A\tau}BI\delta(\tau-\tau')B^\mathsf{T}e^{A^\mathsf{T}\tau'} \\
&= \int_0^t d\tau\, e^{A\tau}BB^\mathsf{T}e^{A^\mathsf{T}\tau}
\end{aligned}$$

(b)

$$\begin{aligned}
\int_0^t d\tau\, \frac{d}{d\tau}e^{A\tau}BB^\mathsf{T}e^{A^\mathsf{T}\tau} &= \int_0^t d\tau\, \left(Ae^{A\tau}BB^\mathsf{T}e^{A^\mathsf{T}\tau} + e^{A\tau}BB^\mathsf{T}e^{A^\mathsf{T}\tau}A^\mathsf{T}\right) \\
&= AW(t) + W(t)A^\mathsf{T}
\end{aligned}$$

(c) We have

$$\begin{aligned}
AW(\infty) + W(\infty)A^\mathsf{T} &= \int_0^\infty d\tau\, \frac{d}{d\tau}e^{A\tau}BB^\mathsf{T}e^{A^\mathsf{T}\tau} \\
&= e^{A\tau}BB^\mathsf{T}e^{A^\mathsf{T}\tau}\Big|_0^\infty \\
&= e^{A\infty}BB^\mathsf{T}e^{A^\mathsf{T}\infty} - BB^\mathsf{T} \\
&= -BB^\mathsf{T}
\end{aligned}$$

provided A is stable (all eigenvalues are negative).

(d) If we consider $\mathbf{u}$ to be a control input, $W(\infty)$ measures the degree to which this input can influence the network state. In particular, if it has any zero eigenvalues then the corresponding dimensions of network space cannot be influenced by the control: they are "uncontrollable".

4 Learning

1. Consider linear regression in N dimensions,

$$y = \mathbf{w} \cdot \mathbf{x} \tag{1}$$

where both $\mathbf{w}$ and $\mathbf{x}$ are vectors in $\mathcal{R}^N$. The weights are trained with P training examples,

$$\{y_\mu, \mathbf{x}_\mu\}, \mu = 1, ..., P \tag{2}$$

and the y_μ are given by the true model,

$$y_\mu = \mathbf{w}^* \cdot \mathbf{x}_\mu \tag{3}$$

with $\mathbf{w}^*$ drawn uniformly from the unit sphere.

Assume you're minimizing the squared error, learning the minimum norm solution, and $\mathbf{x}$ is white,

$$\mathbf{x} \sim \mathcal{N}(0, \mathbf{I})\,. \tag{4}$$

Show that the average generalization error, denoted $\mathcal{E}_g$ (averaged over different training sets), is given by

$$\mathcal{E}_g \equiv \left\langle \left((\widehat{\mathbf{w}} - \mathbf{w}^*) \cdot \mathbf{x} \right)^2 \right\rangle_{P(\mathbf{x})} = \max(0, 1 - P/N)\,. \tag{5}$$

Solution

The generalization error is given by

$$\mathcal{E}_g = (\widehat{\mathbf{w}} - \mathbf{w}^*) \cdot \langle \mathbf{x}\mathbf{x} \rangle_{P(\mathbf{x})} \cdot (\widehat{\mathbf{w}} - \mathbf{w}^*) = |\widehat{\mathbf{w}} - \mathbf{w}^*|^2 \tag{6}$$

where the second equality follows because $\mathbf{x}$ is white. Because there's no model mismatch or noise, for $P < N$ we can fit the training data perfectly,

$$\mathbf{w} \cdot \mathbf{x}_\mu = \mathbf{w}^* \cdot \mathbf{x}_\mu\,. \tag{7}$$

Generically, the $\mathbf{x}_\mu$ span a P-dimensional space. That means you learn P components of $\mathbf{w}^*$. Because we're learning the minimum norm solution, the other $N - P$ components are zero. This means that in an appropriately rotated frame, the least squares minimum norm solution is

$$\widehat{\mathbf{w}} = (w_1^*, w_2^*, ..., w_P^*, 0, ...0)\,. \tag{8}$$

Consequently, the generalization error, Eq. (6), is given by

$$\mathcal{E}_g = \sum_{n=P+1}^{N} \langle (w_i^*)^2 \rangle_{\mathbf{w}^*} \tag{9}$$

Since $\mathbf{w}^*$ is drawn from the unit sphere, the right hand side is $(N - P)/N = 1 - P/N$.

If $P > N$, on the other hand, $\widetilde{\mathbf{w}} = \mathbf{w}^*$ (again because there's no model mismatch or noise), and the generalization error is zero.

2. Consider a neuron whose firing rate, y, is linear in synaptic input, $\mathbf{x}$,

$$y = \mathbf{w} \cdot \mathbf{x}\,. \tag{10}$$

Assume that the update rule for the weight is

$$\tau \frac{d\mathbf{w}}{dt} = -\langle (y - \gamma y^2)\mathbf{x} \rangle_{P(\mathbf{x})} \tag{11}$$

with $\gamma > 0$. The angle brackets indicate an average over $\mathbf{x}$, whose elements x_i are drawn iid from a **non-negative** distribution.

(a) Show that if the weight vector, $\mathbf{w}$, is such that $\gamma y^2 \ll y$ the weights will decay to zero, and if $\gamma y^2 \gg y$ they will grow forever.

(b) There's a fixed point at $y = \gamma^{-1}$, but the previous question suggests that it's unstable. Suppose γ itself is a dynamical variable that evolves according to

$$\frac{d\gamma}{dt} = f(y, \gamma)\,. \tag{12}$$

Choose the function $f(y, \gamma)$ to stabilize the fixed point, and explain why it works. Your explanation can be qualitative or quantitative.

Solution

(a) First consider the case $\gamma y^2 \ll y$. In that case we can neglect the second term in Eq. (11), and write

$$\tau \frac{d\mathbf{w}}{dt} = -\mathbf{w} \cdot \langle \mathbf{x}\mathbf{x} \rangle_{P(\mathbf{x})}\,. \tag{13}$$

The matrix $\langle \mathbf{x}\mathbf{x} \rangle_{P(\mathbf{x})}$ is a covariance matrix, so it has, in the typical case, all positive eigenvalues. Consequently, $\mathbf{w}$ will decay exponentially to zero.

If, on the other hand, $\gamma y^2 \gg y$, we can neglect the first term in Eq. (11), and write

$$\tau \frac{d\mathbf{w}}{dt} = \gamma \langle y^2 \mathbf{x} \rangle_{P(\mathbf{x})}\,. \tag{14}$$

Because the x_i are non-negative, the right hand side is positive, which means $\mathbf{w}$ will increase forever (excluding the case $\langle x_i \rangle = 0$).

(b) To stabilize the fixed point at $y = \gamma y^2$, we want γ to decrease if $y > \gamma^{-1}$, increase if $y < \gamma^{-1}$, and not change if they're equal. For instance,

$$\tau_\gamma \frac{d\gamma}{dt} = \gamma^{-1} - y\,. \tag{15}$$

Stability analysis is difficult, because $\mathbf{w}$ is a vector, but it's clear that γ should change rapidly around the fixed point. Thus, we want τ_γ small compared to τ.

3. Suppose we wanted to derive a learning rule that preserved the sign of the weights, so that Dale's law would remain satisfied during learning. A simple one is a multiplicative rule. As usual, we'll consider the highly over-simplified setting in which a single neuron, y, receives input from neurons, $\mathbf{x} = (x_1, ...)$, mediated by weights, $\mathbf{w} = (w_1, ...)$,

$$y = \mathbf{w} \cdot \mathbf{x}\,. \tag{16}$$

Consider the following weight update rule in discrete time,

$$w_i(t+1) - w_i(t) = \eta w_i(t) y x_i\,. \tag{17}$$

(a) Assume the magnitude of both the input and output are, are bounded: $|x_i| \le x_{\max}$ and $|y| \le y_{\max}$. Write down a condition on the learning rate, η, that will ensure that the weights never change sign.

(b) Add synaptic scaling to the learning rule: after each update,

$$w_i(t+1) \to \frac{w_i(t+1)}{\gamma} \tag{18}$$

where γ is chosen so that after the scaling,

$$\sum_i w_i(t+1) = w_0\,. \tag{19}$$

Given this setup, write down a **local** learning rule in the limit $\eta \to 0$.

Solution

(a) The learning rule, Eq. (17), can be written

$$w_i(t+1) = w_i(t)(1+\eta y x_i)\,. \tag{20}$$

There's no sign change if $\eta|y||x_i| \leq 1$. Given the bounds on x_i and y, this is guaranteed if $\eta \leq 1/(y_{\max}x_{\max})$.

(b) We can combine the two-step learning rule into one step,

$$w_i(t+1) = \frac{w_i(t)(1+\eta y x_i)}{\gamma}\,. \tag{21}$$

Summing both sides over i and using Eqs. (16) and (19), we see that γ must satisfy

$$w_0 = \frac{w_0 + \eta y^2}{\gamma}\,, \tag{22}$$

which implies that

$$\gamma = 1 + \frac{\eta y^2}{w_0}\,. \tag{23}$$

Our learning rule thus becomes

$$w_i(t+1) = \frac{w_i(t)(1+\eta y x_i)}{1+\eta y^2/w_0} = w_i(t)\left(1+\eta y\left(x_i - \frac{y}{w_0}\right)\right) \tag{24}$$

where the last equality is valid in the limit $\eta \to 0$. This can also be written

$$w_i(t+1) - w_i(t) = \eta y w_i\left(x_i - \frac{y}{w_0}\right)\,. \tag{25}$$

As promised, the sum over i of both the left and right side is zero. Finally, we can set $t+1$ to $t+dt$ and η to dt, giving us

$$\frac{dw_i}{dt} = y w_i\left(x_i - \frac{y}{w_0}\right)\,. \tag{26}$$

4. Consider an archer who is trying to hit a 1-dimensional **hidden** target. On each trial, the archer aims at a point, $\mu \in \mathcal{R}$. The archer isn't perfect, so the arrow hits a point $y = \mu + z$ where $z \sim \mathcal{N}(0, \sigma^2)$ is Gaussian noise. The archer observes y and gets a reward R (which in general will be a function of y and the hidden target). We treat each trial as a new episode, so that $J = \mathbb{E}[R]$.

Assume that the parameter μ is learned using a REINFORCE learning rule: after each trial, the policy parameter(s) are updated based on the (single sample) Monte-Carlo estimate for the Policy Gradient.

(a) Write down the update rule for μ, for arbitrary reward R. Hint: the policy must be stochastic.

(b) Assume the real target (which is hidden from the archer) is located at m, and the reward function is $R = -(y-m)^2$. What is the expected change in μ per trial?

Solution

(a) The update rule for REINFORCE with learning rate η is

$$\Delta\mu = \eta R\,\frac{\partial \log \pi(y)}{\partial \mu}\,. \tag{27}$$

A convenient policy is a Gaussian one (although you could use anything you want),

$$\pi(y) = \frac{e^{-(y-\mu)^2/2\sigma_0^2}}{\sqrt{2\pi\sigma_0^2}}\,. \tag{28}$$

The derivative is easy, and we have

$$\Delta\mu = \frac{\eta R}{\sigma_0^2}\,(y-\mu)\,. \tag{29}$$

This is quite reasonable: increase μ if it's smaller than y; decrease it if it's larger than y.

(b) Inserting $R = -(y-m)^2$ into the above expression gives us

$$\Delta\mu = -\frac{\eta}{\sigma_0^2}\,(y-m)^2(y-\mu)\,. \tag{30}$$

To compute the expected values, let $y = \mu + z$ and average over z. Because z is Gaussian, we need only the quadratic and zeroth order terms (except that there is no zeroth order term),

$$(y-m)^2(y-\mu) = (\mu+z-m)^2 z = 2z^2(\mu-m) + \text{odd powers of } z\,. \tag{31}$$

Thus,

$$\mathbb{E}[\Delta\mu] = -\frac{2\eta\sigma^2}{\sigma_0^2}\,(\mu-m)\,. \tag{32}$$

Alternatively, we could use the policy gradient theorem, which states that the update rule is a stochastic estimate of the gradient of J . In our case, J is simply $\mathbb{E}[R]$, so we write

$$\mathbb{E}[\Delta\mu] = \eta\frac{\partial\mathbb{E}[R]}{\partial\mu}\,. \tag{33}$$

The expected reward is, then,

$$\mathbb{E}[R] = \mathbb{E}[-(y-m)^2] = \mathbb{E}[-(\mu+z-m)^2] = -\sigma^2 - (\mu-m)^2\,. \tag{34}$$

Inserting this into Eq. (33) and taking the derivative, we recover Eq. (32) – up to a scale factor. The fact that we're off by a scale factor is fine, since the learning rate is arbitrary.