Review Solutions, Exam 1

1. Our definition of learning: Learning is the process of building a desirable association between stimulus and response, and is measured through behavior on stimulus not seen before. This differs significantly from the dictionary's definition- The dictionary allows for memorization, which we do not (by measuring learning on unseen stimulus), and implies that there exists a measure of desirablity.

2. Superstitious behavior is behavior that results from an incorrect association between stimulus and response- that is, in reality there was no association. This would fit in the dictionary's definition of learning, but not in ours.

3. The outcome was that, in general, people will tend to match the true probabilities for the lights- if signal light 1 comes on 75% of the time, people will tend to choose that light 75% of the time. This is similar to the card experiment (choose whether A or B is coming up next), and we saw that we did basically follow the probabilities.

4. While this does not maximize the reward (to maximize the reward, we should always choose the light or card with the highest probability), it is desirable in the sense that in the real world, such probabilities will not remain constant.

5. In the $N-$armed bandit problem, we are given N slot machines (or N arms on a single slot), with N different payoffs. The goal is to try to maximize our payoff.

6. The greedy algorithm stated that we should always choose the slot (or arm) with the highest estimated payoff, where ties are broken at random. The $\epsilon-$greedy algorithm stated that, with probability ϵ, we should choose a machine at random. The $\epsilon-$greedy algorithm is better initially since our estimates might be incorrect- the greedy algorithm may get stuck on a suboptimal machine.

7. The softmax strategy was to construct probabilities from the estimated payoffs. That was, if $Q_t(a)$ is the estimated payoff at time t for machine a, then:

$$\pi_t(a) = \frac{\exp\left(\frac{Q_t(a)}{\tau}\right)}{\sum_{k=1}^{N} \exp\left(\frac{Q_t(k)}{\tau}\right)}$$

   where $\pi_t(a)$ is the probability of selecting slot a at time t. We should think of τ as a way of controlling the update- if τ is very large, this algorithm will choose machines at random since all probabilities will be approximately equal. If τ is small, it behaves like the greedy algorithm.

   **NOTE THE ERROR, p. 37: The limits being computed (under the second item after (2.) should be as $\tau \to 0$, not as $\tau \to \infty$.

8. The pursuit strategy, similar to the "Win-Stay, Lose-Shift" strategy is that, when we win, increase the probability of that occurring again. If we lose, decrease the probability of that occurring again. In particular, given a set of estimates of the payoffs, make the probability of choosing the best machine greater, and decrease the others. The change in probabilities are proportional to their distance (to 1 or 0):

$$\begin{aligned} \pi_{t+1}(a) &= \pi_t(a) + \beta(1 - \pi_t(a)) \quad \text{for the winner} \\ \pi_{t+1}(a) &= \pi_t(a) + \beta(0 - \pi_t(a)) \quad \text{for the losers} \end{aligned}$$

9. Hebb's idea was that, if neuron A repeatedly takes part in firing neuron B, then some process (growth or metabolic) takes place so that its easier in the future for cell A to fire cell B.

10. A linear network implements a linear (or affine) map, $\boldsymbol{y} = W\boldsymbol{x} + \boldsymbol{b}$, where W was the weight matrix, and $\boldsymbol{b}$ is a vector of biases (or resting states).

11. The three update rules:

    - Hebb's Rule: $W^{\text{new}} = W^{\text{old}} + \alpha \boldsymbol{y}\boldsymbol{x}^T$, where $\boldsymbol{y} = W\boldsymbol{x} + \boldsymbol{b}$.
    - Modified Hebb's Rule: $W^{\text{new}} = W^{\text{old}} + \alpha \boldsymbol{t}\boldsymbol{x}^T$, where $\boldsymbol{t}$ is the desired target for $\boldsymbol{x}$.
    - Widrow-Hoff Rule: $W^{\text{new}} = W^{\text{old}} + \alpha(\boldsymbol{t} - \boldsymbol{y})\boldsymbol{x}^T$

12. Matlab Questions:

    (a) A script file is a text file with a series of Matlab commands typed in. A function is like a subroutine that is called from Matlab- a file that defines a function begins with the word `function`.

    (b) This code fragment finds the maximum elements of Q, which is the value 3, appearing at index 2, 5. Therefore, the vector `idx` will be $[2, 5]$.

    (c) The difference is that `rand` uses a uniform distribution of random numbers between 0 and 1, while `randn` uses a normal (Gaussian) distribution with mean 0 and standard deviation 1.

    (d) The command `cumsum` is for cumulative sums. P will be a vector of partial sums of x, $P = [0.3, 0.4, 0.6, 1.0]$.

    (e)

13. We said that the Fourier transform breaks up a complex waveform into a sum of simple waves. How did it do that? Why does this work? (Give a general description in the continuous case). The Fourier transform projects a function into sines and cosines of differing periods. This works because the sines and cosines form an orthogonal basis (this is really a change of coordinates transformation!).

14. In the continuous setting, the Fourier expansion is:

$$f(x) = a_0 + \sum_{k=1}^{\infty} a_k \cos(kx) + \sum_{k=1}^{\infty} b_k \sin(kx)$$

The coefficient a_0 is computed as the projection of f onto the function 1,

$$\frac{1}{2\pi} \int_0^{2\pi} f(x)\, dx$$

(Notice that $2\pi = \|1\|^2$, which is also acceptable) and is also the mean of f. The coefficients a_k are computed as the projection of f onto $\cos(kx)$,

$$\frac{1}{\pi} \int_0^{2\pi} f(x) \cos(kx)\, dx$$

(again, you could also use $\|\cos(kx)\|^2$ in place of π) and the coefficients b_k are the projection of f onto $\sin(kx)$:

$$\frac{1}{\pi} \int_0^{2\pi} f(x) \sin(kx)\, dx$$

15. The inner product is (use integration by parts):

$$\int_0^{2\pi} x\mathrm{e}^x\, dx = x\mathrm{e}^x - \mathrm{e}^x|_0^{2\pi} = \mathrm{e}^{2\pi}(2\pi - 1) + 1$$

16. The Fourier coefficients:

$$a_0 = \frac{1}{\pi} \int_0^{2\pi} x^2\, dx = \frac{8\pi^2}{3}$$

We compute a_3 and b_3 together:

$$\frac{1}{\pi} \int_0^{2\pi} x^2 \mathrm{e}^{3xi}\, dx = \frac{1}{\pi} \mathrm{e}^{3xi} \left(\frac{x^2}{3i} + \frac{2x}{9} - \frac{2}{27i} \right) \Bigg|_0^{2\pi} = -\frac{4\pi}{3} i + \frac{4}{9}$$

so $a_3 = \frac{4}{9}$, and $b_3 = -\frac{4\pi}{3}$

17. Use the relationships written in the Summary section:

$$a_0 = \frac{F(1)}{N} = \frac{32}{64} = \frac{1}{2}$$

The number $F(3)$ corresponds to coefficients a_2 and b_2:

$$a_2 = \frac{2}{N} \mathrm{Real}(F(3)) = 0, b_2 = \frac{-2}{N} \mathrm{Imag}(F(3)) = \frac{-2}{(-64)} 64 = 2$$

The number $F(6)$ corresponds to coefficients a_5 and b_5 in a similar way:

$$a_5 = \frac{2}{N} \mathrm{Real}(F(6)) = \frac{2}{64} 96 = 3, b_5 = 0$$

so $f(x) = \frac{1}{2} + 2\sin(2x) + 3\cos(5x)$.

18. The smallest period is twice the sampling distance (Δx).

19. The first vector is a vector with all ones. The second vector contains the 4 roots of unity, $1, i, -1, -i$. The third vector is: And the fourth vector is:

20. $(3-i)(1+i) + (2+i)(1+i) - 3i = (1+i)(5) - 3i = 5 + 2i$

21. The power spectrum plots the size of each element of F, which are complex. That is, we see the plot of k versus $|F(k)|$.

22. The mean is 4.5, the variance is $41/4$

23. You could compute the covariance using the definition, or see that the second data set is 3 times the first (and both have zero mean). Therefore, $\mathrm{Cov}(x, 3x) = 3\mathrm{Cov}(x, x) = 3 \cdot \frac{10}{5} = 6$.

24. Formally, the definition of the correlation coefficient is:

$$\rho_{xy} = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y}$$

    If we think of the data in x and y as vectors with mean zero, then the correlation coefficient is $\cos(\theta)$, where θ is the angle between them.

25. The covariance matrix for the $p \times n$ matrix X consisting of p points in $\mathbb{R}^n$ is $C = \frac{1}{p} X m^T X m$ where Xm is the mean subtracted matrix (mean is taken over p points).

26. The (i, j)th term of the covariance matrix represents the covariance between the ith and jth columns of X.

27. The linearization is given by:

$$L(x, y) = \begin{bmatrix} -1 \\ 3 \end{bmatrix} + \begin{bmatrix} 3 & 0 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x-3 \\ y-1 \end{bmatrix}$$

    so the approximation is:

$$\begin{bmatrix} -1 \\ 3 \end{bmatrix} + \begin{bmatrix} 3 & 0 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 2-3 \\ 2-1 \end{bmatrix} = \begin{bmatrix} -4 \\ 5 \end{bmatrix}$$

28. To linearize, compute the derivative:

$$\boldsymbol{f}'(t) = \begin{bmatrix} -\pi \sin(\pi t) \\ 2t + 2 \end{bmatrix}$$

    so at $t = 1$, $\boldsymbol{f}(1) = [-1, 3]^T$, $\boldsymbol{f}'(1) = [0, 4]^T$. Putting these together, the linearization is the line:

$$L(t) = \begin{bmatrix} -1 \\ 3 \end{bmatrix} + \begin{bmatrix} 0 \\ 4 \end{bmatrix} (t-1)$$

29. (a) Compute the gradient:

$$\nabla f = [6x + y + 3, x - 2y - 5]$$

so at $(1, 1)$, $\nabla f = [10, -6]$. Therefore, the direction of fastest increase is in the direction $[10, -6]^T$.

(b) The Hessian is the matrix of second partials:

$$Hf = \begin{bmatrix} 6 & 1 \\ 1 & -2 \end{bmatrix}$$

(c) The choice of A is not unique. We could write it using the Hessian- In that case,

$$f(x, y) = \frac{1}{2}\boldsymbol{x}^T \begin{bmatrix} 6 & 1 \\ 1 & -2 \end{bmatrix} \boldsymbol{x} + [3, -5]\boldsymbol{x}$$

(d) The stationary point is where the gradient is zero- We could either use $A\boldsymbol{x} + \boldsymbol{b}$ from the last part, or just rewrite the equations:

$$\nabla f(\boldsymbol{x}) = 0 \Rightarrow \begin{bmatrix} 6 & 1 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -3 \\ 5 \end{bmatrix}$$

$$\begin{bmatrix} x \\ y \end{bmatrix} = \frac{-1}{13} \begin{bmatrix} -2 & -1 \\ -1 & 6 \end{bmatrix} \begin{bmatrix} -3 \\ 5 \end{bmatrix} = \frac{-1}{13} \begin{bmatrix} 1 \\ 33 \end{bmatrix}$$

(e) If we consider the original version of f, we see that f contains neither a (global) maximum or minimum (the coefficient in front of x^2 is positive, in front of y^2 is negative).

30. Notice that the vectors given are orthogonal. Therefore,

$$\alpha_1 = \frac{(1 \cdot 1) + (0 \cdot 3) + (2 \cdot (-2))}{1^2 + 3^2 + 2^2} = \frac{-3}{14}$$

$$\alpha_2 = \frac{(1 \cdot 3) + (0 \cdot (-1)) + (2 \cdot 0)}{3^2 + 1^2} = \frac{3}{10}$$

So, the projection of $\boldsymbol{x}$ onto the plane is:

$$\boldsymbol{z} = \frac{-3}{14} \begin{bmatrix} 1 \\ 3 \\ -2 \end{bmatrix} + \frac{3}{10} \begin{bmatrix} 3 \\ -1 \\ 0 \end{bmatrix} = \frac{1}{35} \begin{bmatrix} 24 \\ -33 \\ 15 \end{bmatrix}$$

The distance from $\boldsymbol{x}$ to the plane is $\|x - z\|$, which is the norm of:

$$\frac{1}{35} \begin{bmatrix} 11 \\ 33 \\ 55 \end{bmatrix}$$

which is about 1.86

31.

$$\boldsymbol{x} = 3 \begin{bmatrix} 6 \\ 1 \end{bmatrix} + (-1) \begin{bmatrix} 1 \\ -2 \end{bmatrix} = \begin{bmatrix} 18 \\ 9 \end{bmatrix}$$

32. The vectors are not orthogonal, so we have to invert the matrix:

$$[x]_{\mathcal{B}} = \frac{-1}{13} \begin{bmatrix} -2 & -1 \\ -1 & 6 \end{bmatrix} \begin{bmatrix} 3 \\ -1 \end{bmatrix} = \frac{1}{13} \begin{bmatrix} 5 \\ 9 \end{bmatrix}$$

33.

$$A = \frac{\boldsymbol{a}\boldsymbol{a}^T}{\boldsymbol{a}^T\boldsymbol{a}} = \frac{1}{10} \begin{bmatrix} 1 & 3 \\ 3 & 9 \end{bmatrix}$$