1. [95 pts] You are given the task of estimating the IQ of each person among a group of subjects. Denote the IQ of each subject by $\theta_i, i = 1, \ldots, S$, where i is denotes the $i-th$ subject, and S is the total number of subjects. You conduct a test where the result of the test for subject i is $x_i, i = 1, \ldots, S$. Suppose that the test results are related to the actual IQ as

$$x_i = \theta_i + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1),$$

and are *independent* for each subject, with identically distributed noise.

(a) [10 pts] Given a dataset of measurements $\mathcal{D} = \{x_1, x_2, \ldots, x_S\}$, specify the likelihood of the data, $p(\mathcal{D}|\theta_1, \ldots, \theta_S)$, then find the Maximum Likelihood estimates $\theta_{ML,i}|\mathcal{D}, i = 1, \ldots, S$.

$$P(D|\theta_1, \ldots \theta_S) = P(x_1, x_2 \ldots x_S | \theta_1, \theta_2 \ldots \theta_S)$$
$$= P(x_1|\theta_1, \ldots \theta_S)\, P(x_2|\theta_1 \ldots \theta_S) \ldots P(x_S|\theta_1 \ldots \theta_S)$$
$$= P(x_1|\theta_1)\, P(x_2|\theta_2) \ldots P(x_S|\theta_S)$$

$$P(D|\theta_1, \ldots \theta_S) = \prod_i P(x_i|\theta_i)$$

$$P(x_i|\theta_i) \sim \mathcal{N}(\theta_i, 1) \quad (\text{as } x_i = \theta_i + \epsilon)$$

$$P(D|\theta_1 \ldots \theta_S) = \prod_i \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i - \theta_i)^2}{2}}$$

$$LL = \log P(D|\theta_1 \ldots \theta_S) = -\sum_{i=1}^{S} \frac{1}{2}(x_i - \theta_i)^2 + \log\left(\frac{1}{\sqrt{2\pi}}\right)$$

$$\frac{\partial LL}{\partial \theta_i} = 0 \Rightarrow \frac{1}{2} \times -2 (x_i - \theta_i)(-1) = 0$$

$$\Rightarrow \theta_i = x_i$$

$$\boxed{\theta_{ML,i|D} = x_i}, \quad i = 1, \ldots S$$

(b) [5 pts] Prove whether the *estimates* of each person's IQ, $\theta_{ML,i}|\mathcal{D}$, are dependent or independent of the results of the test for other people, $x_j, j \neq i$?

From part 'a'

$$\theta_{ML,i|D} = x_i$$

This expression doesn't depend on other x_j, $j \neq i$

$\Rightarrow$ Independent of the results for other people

(c) [5 pts] Given what you have learned from the data $\mathcal{D}$, what is the most likely test result of each person's next IQ test?

$x_i = \theta_i + \epsilon \Rightarrow E[x_i] = E[\theta_i] + E[\epsilon]$ (→ '0')

most likely test result of i^{th} person $= E[x_i] = E[\theta_i]$

$= x_i$ (we have only one sample for i^{th} person)

(d) [20 pts] Now assume that, even though each subject has his or her individual IQ coefficient θ_i, the coefficients of all people do not deviate too much. In other words, we can assume that apriori

$$\theta_i \sim \mathcal{N}(\mu, \tau^2),$$

for some unknown, but fixed, μ and τ. Find the posterior estimates of each person's IQ, $p(\theta_i|\mathcal{D}, \mu, \tau), i = 1, \ldots, S$. Specify the posterior mean $\mathbb{E}[\theta_i|\mathcal{D}, \mu, \tau]$ and the posterior variance $\mathbb{V}[\theta_i|\mathcal{D}, \mu, \tau]$.

$$P(\theta_i \mid D, \mu, \tau) \propto (\text{Prior})(\text{likelihood})$$

$$= P(\theta_i \mid \mu, \tau) \times P(D \mid \theta_i, \mu, \tau)$$

$$P(D \mid \theta_i, \mu, \tau) = P(x_i \mid \theta_i, \mu, \tau) \times P(x_1, x_2 \ldots x_{i-1} x_{i+1} x_S \mid \mu, \tau)$$

$$P(D \mid \theta_i, \mu, \tau) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i - \theta_i)^2}{2}} \times \prod_{j \neq i} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_j - \mu)^2}{2(\tau^2+1)}}$$

$$P(\theta_i \mid D, \mu, \tau) = \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{(\theta_i - \mu)^2}{2\tau^2}} \times \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i - \theta_i)^2}{2}} \times \underbrace{\prod_{j \neq i} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_j - \mu)^2}{2(\tau^2+1)}}}_{(K)}$$

this part doesn't depend on θ_i

$$P(\theta_i \mid D, \mu, \tau) = \frac{K}{2\pi\tau} e^{-\left(\frac{\left[\theta_i^2 - 2\left[\frac{\mu + \tau^2 x_i}{1+\tau^2}\right]\theta_i + \frac{\mu^2 + x_i^2 \tau^2}{1+\tau^2}\right]}{2\frac{\tau^2}{1+\tau^2}} \right)}$$

Continued on next page

Comparing this equation to a Normal distribution. (we can directly compare the expression in the exponent term, as the other terms are just Normalizing factor)

$$\Rightarrow \quad \text{mean} = \frac{\mu + \tau^2 x_i}{1+\tau^2} \quad \left(E[\theta_i | D, \mu, \tau] \right)$$

$$\text{Variance} = \frac{\tau^2}{1+\tau^2} \quad \left(V[\theta_i | D, \mu, \tau] \right).$$

(e) [5 pts] Are the *mean* estimates of each person's IQ, $\mathbb{E}[\theta_i|\mathcal{D},\mu,\tau]$, now dependent or independent of the results of the test of other persons' IQs, $x_j, j \neq i$? Prove your conclusion.

$$E[\theta_i \mid D, \mu, \tau] = \frac{\mu + \tau^2 x_i}{1+\tau^2}$$

Clearly above equation doesn't have x_j terms $(j \neq i)$. (μ, τ are given)

$\therefore$ $E[\theta_i \mid D, \mu, \tau]$ is independent of the results of other person's IQ's

(f) [5 pts] Given what you have learned from the data $\mathcal{D}$ in this second case (1d), what is the most likely test result of each person's next IQ test?

most likely result will be

$$E[x_i / D, \mu, \tau] = E[\theta_i / D, \mu, \tau]$$

$x_i = \theta_i + \epsilon$

$$= \frac{\mu + \tau^2 x_i}{1 + \tau^2}$$

for $i = 1, 2 \ldots S$

(g) [10 pts] Suppose $S = 10$, $\mu = 120$, and $\mathcal{D} = \{111, 113, 115, \ldots, 129\}$. Plot the graph of $\mathbb{E}[\theta_i|\mathcal{D}, \mu, \tau]$ as a function of τ, for $\tau \in [0, 5]$. Plot all estimates for $i = 1, \ldots, S$ in one graph where τ is on the horizontal axis. What happens as $\tau \to 0$ and $\tau \to \infty$?

$$E[\theta_i | D, \mu, \tau] = \frac{\mu + \tau^2 x_i}{1+\tau^2}$$

$$= \frac{\mu + \tau^2(\mu + x_i - \mu)}{1+\tau^2}$$

$$E[\theta_i | D, \mu, \tau] = \mu + \frac{(x_i - \mu)\tau^2}{1+\tau^2}$$

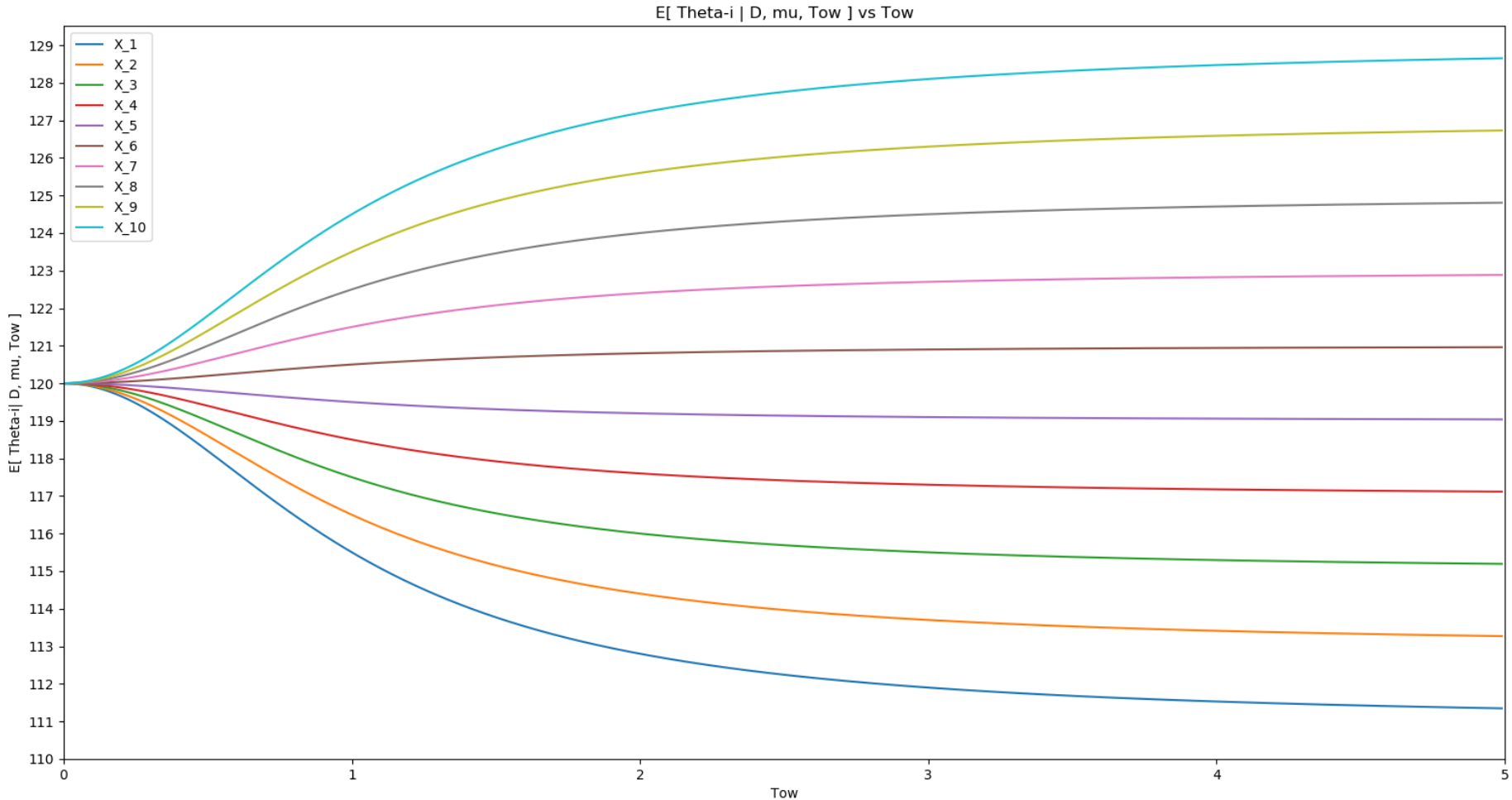
$$\lim_{\tau \to 0} E[\theta_i | D, \mu, \tau] = \mu + \frac{(x_i - \mu)0}{1+0^2} = \mu = 120$$

$$\lim_{\tau \to \infty} E[\theta_i | D, \mu, \tau] = \mu + \frac{(x_i - \mu)}{(1/\tau^2 + 1)} = x_i$$

As $\tau \to \infty$, $E[\theta_i | D, \mu, \tau]$ tends to $\theta_{iML|D}$.

As $\tau \to 0$, $E[\theta_i | D, \mu, \tau]$ tends to 'μ' (mean of apriori of θ_i)

Plots in next page


E[Theta-i | D, mu, Tow] vs Tow
X_1
X_2
X_3
X_4
X_5
X_6
X_7
X_8
X_9
X_10
E[Theta-i| D, mu, Tow]
Tow
129
128
127
126
125
124
123
122
121
120
119
118
117
116
115
114
113
112
111
110
0
1
2
3
4
5

(h) [15 pts] Compute the expression for evidence $p(\mathcal{D}|\mu, \tau)$.

$$P(D|\mu,\tau) = P(x_1, x_2 \ldots x_S | \mu,\tau)$$

$$x_i = \theta_i + \epsilon \longrightarrow \mathcal{N}(0,1)$$
$$\downarrow$$
$$\mathcal{N}(\mu,\tau^2)$$

x_i is sum of two gaussians.

$$\Rightarrow x_i \sim \mathcal{N}(\mu, \tau^2+1)$$

given μ,τ all the x_i's are independent of each other.

$$\Rightarrow P(D|\mu,\tau) = \prod_i P(x_i|\mu,\tau)$$

$$= \prod_i \frac{1}{\sqrt{2\pi(\tau^2+1)}} e^{-\frac{(x_i-\mu)^2}{2(\tau^2+1)}}$$

$$= \left(\frac{1}{\sqrt{2\pi(\tau^2+1)}}\right)^S e^{-\frac{1}{2(\tau^2+1)}\sum_i (x_i-\mu)^2}$$

(i) [10 pts] Given the dataset in (1g), find the estimates of μ and τ, μ^* and τ^* that maximize the evidence.

$$P(D|\mu,\tau) = \prod_i \frac{1}{\sqrt{2\pi(\tau^2+1)}} e^{-\frac{(x_i-\mu)^2}{2(\tau^2+1)}}$$

$$= \left(\frac{1}{\sqrt{2\pi(\tau^2+1)}}\right)^S e^{-\frac{1}{2(\tau^2+1)}\sum_{i=1}^{S}(x_i-\mu)^2}$$

$$L = \log P(D|\mu,\tau) = -\frac{1}{2(\tau^2+1)}\sum_{i=1}^{S}(x_i-\mu)^2 + \frac{S}{2}\log\left(\frac{1}{2\pi(\tau^2+1)}\right)$$

$$\frac{\partial L}{\partial \mu} = 0 \Rightarrow -\frac{1\times 2}{2(\tau^2+1)}\sum_{i=1}^{S}(x_i-\mu)(-1) = 0$$

$$\Rightarrow \mu^* = \frac{1}{S}\sum_{i=1}^{S} x_i = 120 \left(\frac{\overbrace{111+113+\ldots 129}^{10 \text{ terms}}}{10}\right)$$

$$\boxed{\mu^* = 120}$$

$\frac{\partial L}{\partial \tau} = 0$, Substituting given values

$$\frac{1\times 2\tau}{2(\tau^2+1)^2}\sum_{i=1}^{S}(x_i-\mu)^2 - \frac{S}{2}\times\frac{1}{2\pi(\tau^2+1)}\times 2\pi\times 2\tau = 0$$

$$\frac{330\tau}{(\tau^2+1)^2} - \frac{10\times 4\pi\times\tau}{2\times 2\pi\times(\tau^2+1)} = 0$$

$$\Rightarrow \tau(330 - 10(\tau^2+1)) = 0$$

$$\tau^* = \sqrt{32}$$

(j) [10 pts] Now compute the posterior IQ estimates for each subject $\mathbb{E}[\theta_i|\mathcal{D}, \mu^*, \tau^*]$ and $\mathbb{V}[\theta_i|\mathcal{D}, \mu^*, \tau^*]$. How do they differ from the ML estimates of the subjects' IQ?

$$E[\theta_i \mid D, \mu^*, \tau^*] = \frac{\mu^* + (\tau^*)^2 x_i}{1 + (\tau^*)^2}$$

$$= \frac{120 + 32 x_i}{33}$$

$$V[\theta_i \mid D, \mu^*, \tau^*] = \frac{\tau^{*2}}{1 + (\tau^*)^2} = \frac{32}{33}$$

$$E[\theta_i \mid D, \mu^*, \tau^*] - x_i = \frac{120 + 32 x_i}{33} - x_i = \frac{120 - x_i}{33}$$

Calculated values are in next page.

The difference between Likelihood and posterior is the inclusion of prior in posterior case.

i	θML,i|D	E[θi |D, μ*, τ*]	V[θi |D, μ*, τ*]
1	111	111.272727273	0.969697
2	113	113.212121212	0.969697
3	115	115.151515152	0.969697
4	117	117.090909091	0.969697
5	119	119.03030303	0.969697
6	121	120.96969697	0.969697
7	123	122.909090909	0.969697
8	125	124.848484848	0.969697
9	127	126.787878788	0.969697
10	129	128.727272727	0.969697

2. [40 pts] In class we discussed the problem of minimizing the KL divergence between the empirical density $\hat{p}(x|\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N}\delta(x - x_i)$, for some dataset $\mathcal{D} = \{x_i\}_{i=1}^{N}$, and some model-based density $p(x|\theta)$, where θ are the model parameters.

(a) [5 pts] Write down the expression for the KL divergence between the empirical density $\hat{p}(x|\mathcal{D})$ and the Gaussian $p(x|\mu) = \mathcal{N}(x|\mu, \sigma^2 = 1)$, as a function of $\theta = \mu$. Simplify it as much as possible.

$$KL\left(\hat{P}(x|D) \,||\, P(x|\mu)\right) = \int_x \hat{P}(x|D) \log \frac{\hat{P}(x|D)}{P(x|\mu)} dx$$

$$= \sum_{j=1}^{N} \hat{P}(x_j|D) \log \frac{\hat{P}(x_j|D)}{P(x_j|\mu)}$$

$$= \sum_{j=1}^{N} \frac{1}{N} \log \frac{\frac{1}{N}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{(x_j-\mu)^2}{2}}}$$

$$= \sum_{j=1}^{N} \frac{1}{N} \log \frac{\sqrt{2\pi}}{N} \times e^{\frac{(x_j-\mu)^2}{2}}$$

$$= \frac{1}{N}\left[\sum_{j=1}^{N} \frac{(x_j-\mu)^2}{2} + \sum_{j=1}^{N} \log \frac{\sqrt{2\pi}}{N}\right]$$

$$= \frac{1}{N}\left[\sum_{j=1}^{N} \frac{(x_j-\mu)^2}{2} + N \log \frac{\sqrt{2\pi}}{N}\right]$$

$$= \frac{1}{2N} \sum_{j=1}^{N} (x_j-\mu)^2 + \log \frac{\sqrt{2\pi}}{N}$$

(b) [5 pts] What is the estimate of μ that minimizes the KL divergence above? Derive the expression.

$$\frac{\partial \cdot KL(\hat{P} \| P)}{\partial \mu} = 0.$$

$$\Rightarrow \frac{1}{2N}\left[2\sum_{j=1}^{N}(x_j-\mu)(-1)\right]=0$$

$$\Rightarrow \sum_{j=1}^{N}(\mu-x_j)=0$$

$$\Rightarrow \mu = \frac{1}{N}\sum_{j=1}^{N} x_j$$

(c) [25 pts] Write down the expression for the squared Wasserstein $p = 2$ (Euclidean) distance (W_2^2) between the empirical density $\hat{p}(x|\mathcal{D})$ and the Gaussian $p(x|\mu) = \mathcal{N}(x|\mu, \sigma^2 = 1)$.

$$W_2^2 = \inf_{\gamma \in \Gamma(\hat{P}, P)} \int d(x,y)^2 \gamma(x,y)\, dx\, dy$$

$$\gamma(x,y) = \left[\frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2}}\right]\left[\frac{1}{N}\sum_{i=1}^{N} \delta(x - x_i)\right]$$

$$= \int (x-y)^2 \gamma(x,y)\, dx\, dy$$

$$= \frac{1}{\sqrt{2\pi}} \int_x \int_y (x-y)^2 e^{-\frac{(y-\mu)^2}{2}} \left[\frac{1}{N}\sum_{i=1}^{N} \delta(x - x_i)\right] dx\, dy$$

$$= \frac{1}{\sqrt{2\pi} N} \sum_{i=1}^{N} \int_y (x_i - y)^2 e^{-\frac{(y-\mu)^2}{2}}\, dy$$

$$= \frac{1}{N}\sum_{i=1}^{N} \int (y - x_i)^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2}}\, dy$$

$$= \frac{1}{N}\sum_{i=1}^{N} E\left[(y - x_i)^2\right] = \frac{1}{N}\sum_{i=1}^{N} E\left[y^2 - 2x_i y + x_i^2\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N} \left((1+\mu^2) - 2\mu x_i + x_i^2\right)$$

$$= (1+\mu^2) - 2\mu\left(\frac{1}{N}\sum_i x_i\right) + \frac{1}{N}\sum_i x_i^2$$

(d) [5 pts] What is the estimate of μ that minimizes the W_2^2 distance? Derive the expression.

$$\frac{\partial W_2^2}{\partial \mu} = 0$$

$$\Rightarrow \frac{\partial \left((1+\mu^2) - 2\mu\left(\frac{1}{N}\sum x_i\right) + \frac{1}{N}\sum_i x_i^2 \right)}{\partial \mu} = 0$$

$$\Rightarrow 2\mu - 2 \times \frac{1}{N}\sum x_i = 0$$

$$\Rightarrow \mu = \frac{1}{N}\sum_i x_i$$

(a) [10 pts] Sketch a possible decision boundary corresponding to

$$\boldsymbol{w}^* = \arg\max_{\boldsymbol{w}} LL(\boldsymbol{w}).$$

Justify your sketch. Is the answer (decision boundary) unique? What is the classification error on the training set?

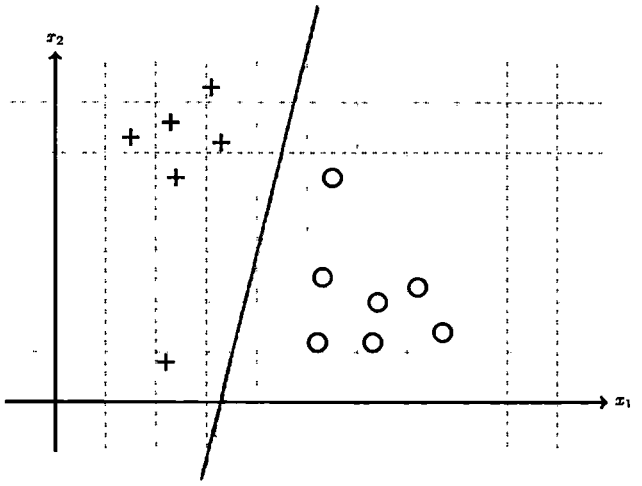


Figure 7: Dataset $\mathcal{D}$ of points from classes "+" and "o".

- **Explanation for choosing this decision boundary**: I choose the decision boundary that separates the two classes with minimum error in the train data and that is close to origin. Because high Likelihood means low error and these are the decision boundaries that lead to low error value to the train data.
- **Decision boundary is unique.**
  Explanation: As the likelihood function is convex, we have a global maximum.
  Practical Scenario: In practice we use numeric optimization to estimate the weight parameters (W) of the model. If we do the gradient descent we might end up at different Ws every time. However, they will be very close to Global maximum of Likelihood function. By adjusting the learning rate, we can try to reach the Global maxima.
- **Misclassification error:** Number of misclassification points = 0

(b) [10 pts] Now sketch a possible decision boundary corresponding to

$$\boldsymbol{w}^* = \arg\max_{\boldsymbol{w}} \left(LL(\boldsymbol{w}) - \lambda w_0^2 \right),$$

where we *heavily* regularize on w_0 (i.e., large λ.) Justify your sketch. Is the answer (decision boundary) unique? What is the classification error on the training set?

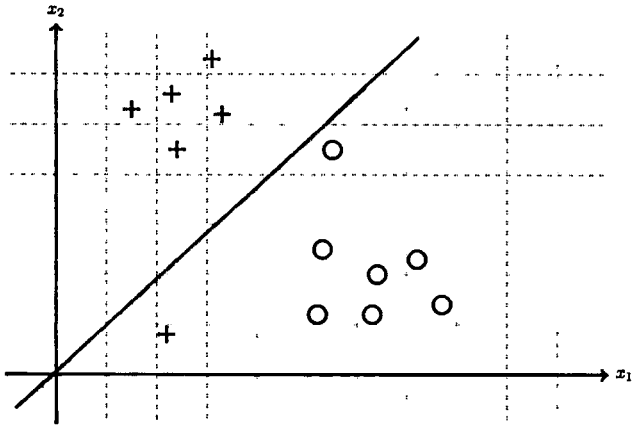


Figure 8: Dataset $\mathcal{D}$ of points from classes "+" and "o".

- Explanation for choosing this decision boundary: Since λ is large, W0 tends to zero as we want to maximize Likelihood function. So, decision boundary will be very close to origin and will be approximately of the form $W_1X_1 + W_2X_2 \approx 0$, as $W0$ tends to zero.

  i choose the decision boundary that separates the two classes with minimum error in the train data and that is close to origin. Because high Likelihood means low error and these are the decision boundaries that lead to low error value to the train data.

- **Decision Boundary is Unique**
  Explanation: As the likelihood is convex, we have a global maximum for the Likelihood. Generally, we use numeric optimization to estimate the weight parameters (W) of the model. So, practically we may not get the same results everytime.
- **Misclassification error:** Number of misclassification points = 1

(c) [10 pts] Next sketch a possible decision boundary corresponding to

$$\boldsymbol{w}^* = \arg\max_{\boldsymbol{w}} \left(LL(\boldsymbol{w}) - \lambda w_1^2 \right),$$

where we *heavily* regularize on w_1. Justify your sketch. What is the classification error on the training set?

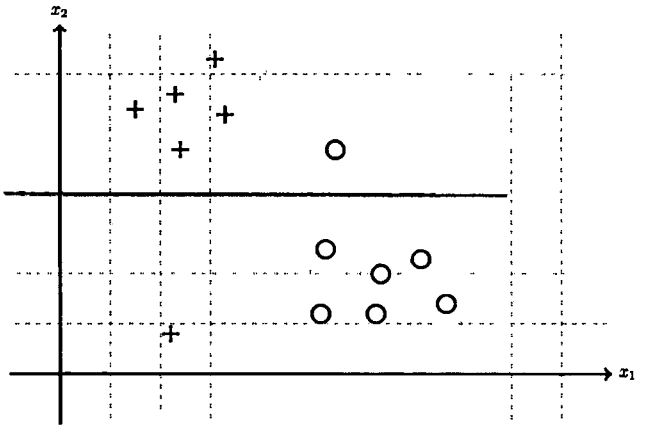


Figure 9: Dataset $\mathcal{D}$ of points from classes "+" and "o".

- **Explanation for choosing this decision boundary:** Since λ is large, W_1 tends to zero as we want to maximize Likelihood function. So, the decision boundary will be nearly parallel to X_1 axis (horizontal axis) and the decision boundary will be of form $W_0 + W_2 X_2 \approx 0$.

  I choose a decision boundary that separates the two classes with minimum error in the train data and nearly parallel to X1 axis. Because high Likelihood means low error and these are the decision boundaries that lead to high likelihood value to the train data.

- **Misclassification error:** Number of data points misclassified = 2.

(d) [10 pts] Finally, sketch a possible decision boundary corresponding to

$$\boldsymbol{w}^* = \arg\max_{\boldsymbol{w}} \left(LL(\boldsymbol{w}) - \lambda w_2^2 \right),$$

where we *heavily* regularize on w_2. Justify your sketch. What is the classification error on the training set?

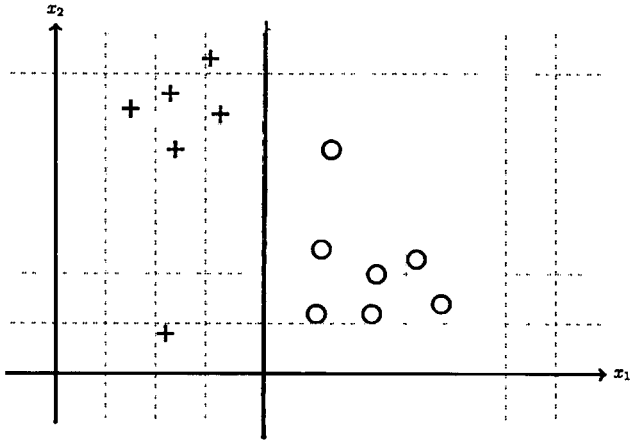


Figure 10: Dataset $\mathcal{D}$ of points from classes "+" and "o".

- **Explanation for choosing this decision boundary:** Since Lambda is large, W_2 tends to zero as we want to maximize Likelihood function. So, the decision boundary will be nearly parallel to X_2 axis (Vertical axis), as the decision boundary will be of form $W_0 + W_1 X_1 \approx 0$. I choose a decision boundary that separates the two classes with minimum error in the train data and parallel to X_2 axis. Because high Likelihood means low error and these are the decision boundaries that lead to high likelihood value to the train data.
- Misclassification error: Number of data points misclassified = 0.

(e) [40 pts] Implementation G_L and experiments

i. Your working files are:
   - Lasso_Regression_Classifier.ipynb
   - cs536_3 → models → Lasso_Regression_Classifier.py
   - feature_extraction.py
   - logistic.py

ii. Please run the feature_extraction.py, then it creates feature set in the feature_data folder.

iii. Run logistic.py then it will give you the logistic classifier accuracies based on the created feature.

iv. Fill out your code as appropriate(Lasso_Regression_Classifier.ipynb and Lasso_Regression_Classifier.py).

v. Find the validation accuracy according to the different L values. Based on the validation accuracy, choose the My L and find a test accuracy based on the My L.

vi. Fill out the table 1. The default value for the average_win was 40. Please modify average_win in the feature_extraction.py and Lasso_Regression_Classifier.ipynb and check validation/test results again in the new feature set.

Table 2: Experiment Results

Average_win	**Logistic Accuracy**	*LandValidationResults*									*TestResults*	
		0.1	**1**	**10**	**13**	**15**	**17**	**21**	**30**	**100**	**My L**	**Test accuracy**
20	0.825	0.3	0.84	0.855	0.86	0.85	0.85	0.845	0.6	0.59	13	0.375
40	0.85	0.3	0.74	0.755	0.61	0.75	0.6	0.745	0.75	0.74	10	0.525
80	0.85	0.45	0.76	0.815	0.815	0.815	0.815	0.735	0.735	0.615	17	0.625
120	0.85	0.45	0.85	0.85	0.85	0.85	0.85	0.8	0.8	0.75	17	0.7
240	0.825	0.475	0.75	0.75	0.75	0.7	0.675	0.675	0.675	0.625	13	0.65