EECE5644 Fall 2025 - Assignment 2

Dawei Wang

October 25, 2025

1 Question 1

1.1 Part 1

The theoretically optimal classifier that minimizes probability of error is the Bayes classifier, which decides class 1 if:

$$P(L = 1|\mathbf{x}) > P(L = 0|\mathbf{x}) \tag{1}$$

Bayes' theorem:

$$\frac{P(L = 1)p(\mathbf{x}|L = 1)}{p(\mathbf{x})} > \frac{P(L = 0)p(\mathbf{x}|L = 0)}{p(\mathbf{x})} \tag{2}$$

Which simplifies to: decide class 1 if $P(L = 1)p(\mathbf{x}|L = 1) > P(L = 0)p(\mathbf{x}|L = 0)$.
Given the class-conditional pdfs:

$$p(\mathbf{x}|L = 0) = 0.5 \cdot g(\mathbf{x}|\mathbf{m}_{01}, \mathbf{C}) + 0.5 \cdot g(\mathbf{x}|\mathbf{m}_{02}, \mathbf{C}) \tag{3}$$

$$p(\mathbf{x}|L = 1) = 0.5 \cdot g(\mathbf{x}|\mathbf{m}_{11}, \mathbf{C}) + 0.5 \cdot g(\mathbf{x}|\mathbf{m}_{12}, \mathbf{C}) \tag{4}$$

Results on Validation Set ($D^{10K}_{validate}$):

Metric	Value
Probability of Error	0.2807
Accuracy	0.7193
True Negatives	5017
False Positives	923
False Negatives	1884
True Positives	2176
ROC AUC	0.7825
Min-P(error) Operating Point	FPR=0.1556, TPR=0.5360

Table 1: Optimal Bayes Classifier Performance

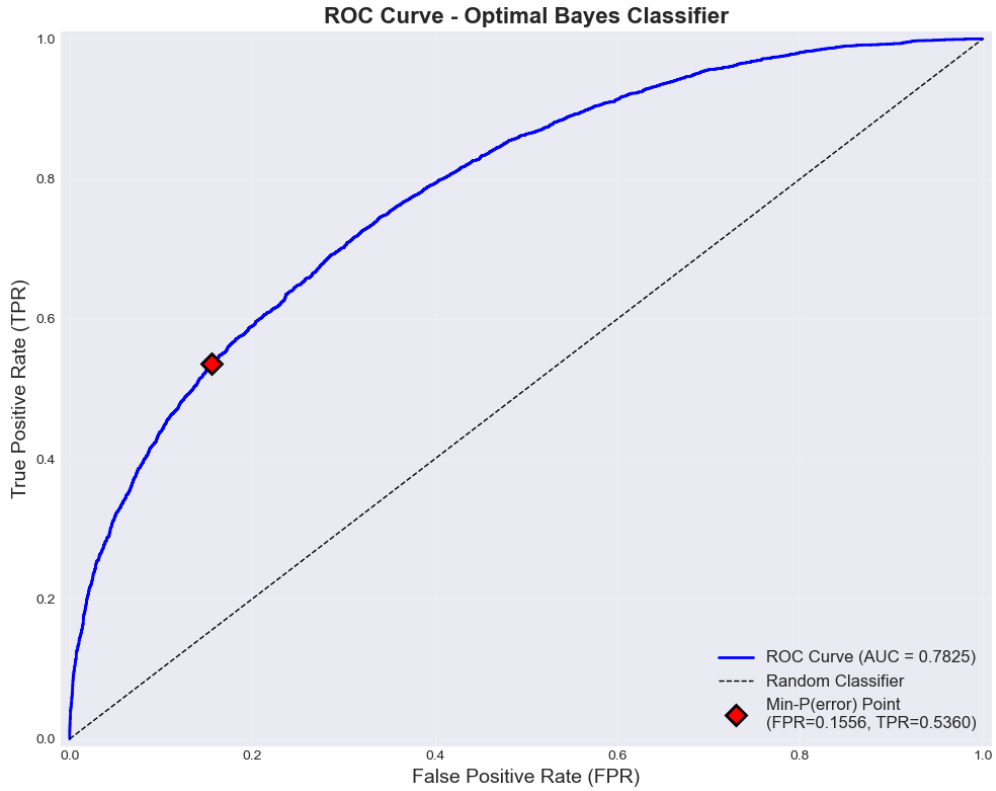

Figure 1: ROC Curve for Optimal Bayes Classifier

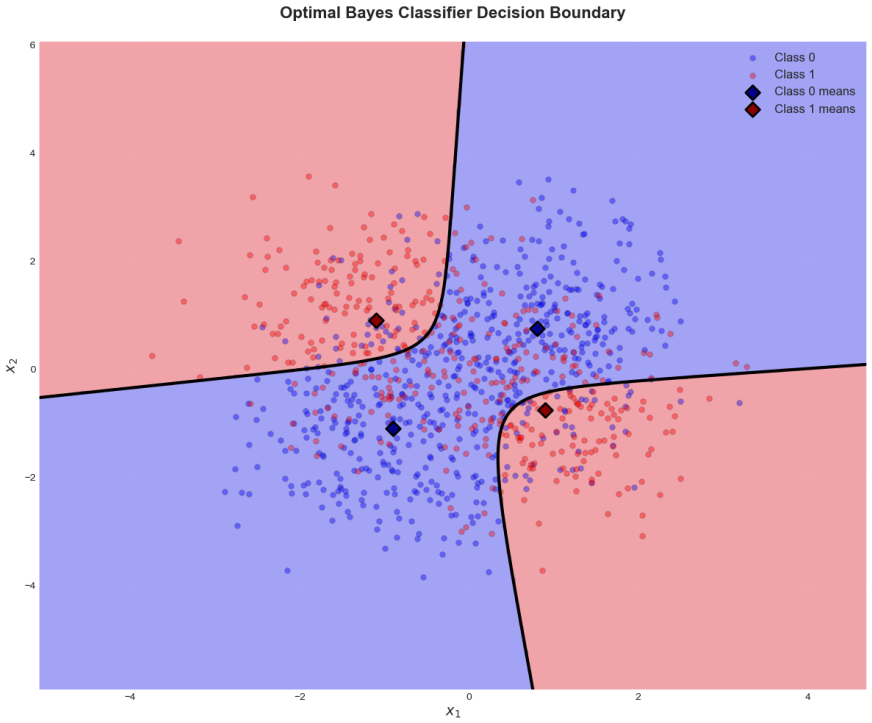

Figure 2: Optimal Bayes Decision Boundary

1.2 Part 2

1.2.1 (a) Logistic-Linear Model

The logistic-linear model approximates the posterior probability as:

$$h(\mathbf{x}, \mathbf{w}) = \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{z}(\mathbf{x})}} \tag{5}$$

where $\mathbf{z}(\mathbf{x}) = [1, x_1, x_2]^T$.

Maximum likelihood estimation minimizes the negative log-likelihood:

$$\mathcal{L}(\mathbf{w}) = -\sum_{i=1}^{N} [y_i \log h(\mathbf{x}_i, \mathbf{w}) + (1 - y_i) \log(1 - h(\mathbf{x}_i, \mathbf{w}))] \tag{6}$$

Results:

Training Set Size	w_0	w_1	w_2
50	-0.627	-0.196	-0.367
500	-0.350	-0.026	0.108
5000	-0.413	-0.030	0.103

Table 2: Logistic-Linear Model Parameters

Training Set Size	Validation Error	Accuracy
50	0.5151	0.4849
500	0.4003	0.5997
5000	0.4047	0.5953

Table 3: Logistic-Linear Model Performance

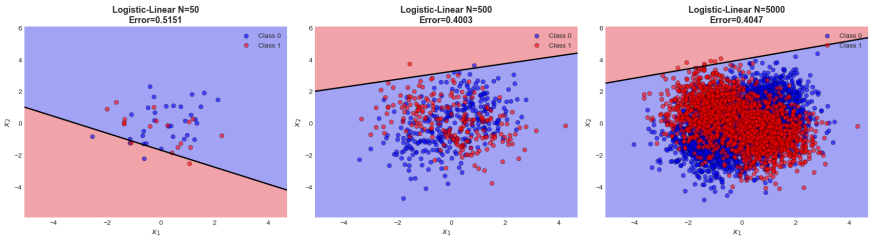

Figure 3: Logistic-Linear Decision Boundaries

1.2.2 Part 2(b): Logistic-Quadratic Model

The logistic-quadratic model uses:

$$\mathbf{z}(\mathbf{x}) = [1, x_1, x_2, x_1^2, x_1 x_2, x_2^2]^T \tag{7}$$

Results:

N	w_0	w_1	w_2	w_3	w_4	w_5
50	-0.707	-0.598	-0.532	0.205	-1.517	-0.343
500	-0.440	-0.026	0.130	0.128	-0.906	-0.018
5000	-0.495	-0.064	0.066	0.077	-0.822	-0.031

Table 4: Logistic-Quadratic Model Parameters

Training Set Size	Validation Error	Accuracy
50	0.2988	0.7012
500	0.2857	0.7143
5000	0.2855	0.7145

Table 5: Logistic-Quadratic Model Performance

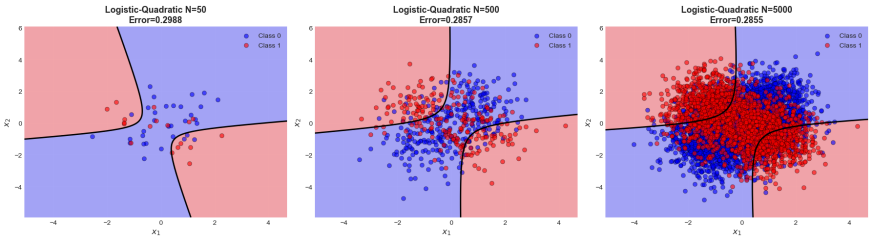

Figure 4: Logistic-Quadratic Decision Boundaries

1.3 Discussion

Model	Error Rate	Accuracy
Optimal Bayes	0.2807	0.7193
Linear (N=50)	0.5151	0.4849
Linear (N=500)	0.4003	0.5997
Linear (N=5000)	0.4047	0.5953
Quadratic (N=50)	0.2988	0.7012
Quadratic (N=500)	0.2857	0.7143
Quadratic (N=5000)	0.2855	0.7145

Table 6: Comprehensive Model Performance Comparison

Figure 5: Error Rate Comparison Across Models and Training Set Sizes

Results:

The quadratic model clearly outperforms the linear model, achieving error rates close to the Bayes optimal (0.2855 vs 0.2807). The linear model's performance is worse (0.4003-0.4047), indicating the decision boundary is inherently nonlinear.

For training set size effects, linear models show minimal improvement beyond 500 samples (0.4003 $\rightarrow$ 0.4047), suggesting the model class is insufficient. Quadratic models improve rapidly from 50 samples (0.2988) to 500 samples (0.2857), then plateau.

With 5000 samples, the quadratic model nearly matches the Bayes optimal performance (within 0.5%), while the linear model remains approximately 12% worse, indicating that model capacity is more critical than sample size when the true decision boundary is complex.

2 Question 2

2.1 Problem Formulation

Given the model $y = c(\mathbf{x}, \mathbf{w}) + v$, where $c(\cdot, \mathbf{w})$ is a cubic polynomial and $v \sim \mathcal{N}(0, \sigma^2)$:

$$c(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_1 x_2 + w_5 x_2^2 + w_6 x_1^3 + w_7 x_1^2 x_2 + w_8 x_1 x_2^2 + w_9 x_2^3 \quad (8)$$

2.2 Maximum Likelihood Estimator

The likelihood function for N independent samples:

$$p(\mathcal{D}|\mathbf{w}, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{\Phi}_i^T \mathbf{w})^2}{2\sigma^2}\right) \quad (9)$$

Taking the negative log-likelihood:

$$-\log p(\mathcal{D}|\mathbf{w}, \sigma^2) = \frac{N}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \mathbf{\Phi}_i^T\mathbf{w})^2 \tag{10}$$

Minimizing with respect to $\mathbf{w}$ yields:

$$\mathbf{w}_{ML} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{y} \tag{11}$$

ML Estimation Results:

Parameter	ML Estimate
w_0	0.3217
w_1	0.1559
w_2	0.0580
w_3	-0.0035
w_4	0.0417
w_5	-0.1980
w_6	-0.0111
w_7	0.0026
w_8	-0.0345
w_9	-0.0703
Training MSE	3.2243
Validation MSE	4.8862

Table 7: Maximum Likelihood Estimation Results

2.3 MAP Estimator

With prior $\mathbf{w} \sim \mathcal{N}(0, \gamma\mathbf{I})$, the posterior is:

$$p(\mathbf{w}|\mathcal{D}, \gamma, \sigma^2) \propto p(\mathcal{D}|\mathbf{w}, \sigma^2)p(\mathbf{w}|\gamma) \tag{12}$$

The MAP estimator maximizes the log-posterior:

$$\mathbf{w}_{MAP} = \arg\min_{\mathbf{w}}\left\{\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{\Phi}\mathbf{w}\|^2 + \frac{1}{2\gamma}\|\mathbf{w}\|^2\right\} \tag{13}$$

This yields:

$$\mathbf{w}_{MAP} = \left(\mathbf{\Phi}^T\mathbf{\Phi} + \frac{\sigma^2}{\gamma}\mathbf{I}\right)^{-1}\mathbf{\Phi}^T\mathbf{y} \tag{14}$$

MAP Estimation Results ($\gamma = 1.0$):

Parameter	MAP	ML
w_0	0.2841	0.3217
w_1	0.1518	0.1559
w_2	0.0542	0.0580
w_3	-0.0032	-0.0035
w_4	0.0408	0.0417
w_5	-0.1818	-0.1980
w_6	-0.0111	-0.0111
w_7	0.0026	0.0026
w_8	-0.0338	-0.0345
w_9	-0.0737	-0.0703
Validation MSE	4.8472	4.8862

Table 8: MAP vs ML Estimation ($\gamma = 1.0$)

2.4 Hyperparameter Analysis

Evaluated MAP estimator for 50 values of γ ranging from 10^{-6} to 10^{6}.

Optimal Results:

- Optimal γ: 5.4287×10^{-6}
- Minimum validation MSE: 4.2685
- ML validation MSE: 4.8862
- Improvement over ML: 12.64%

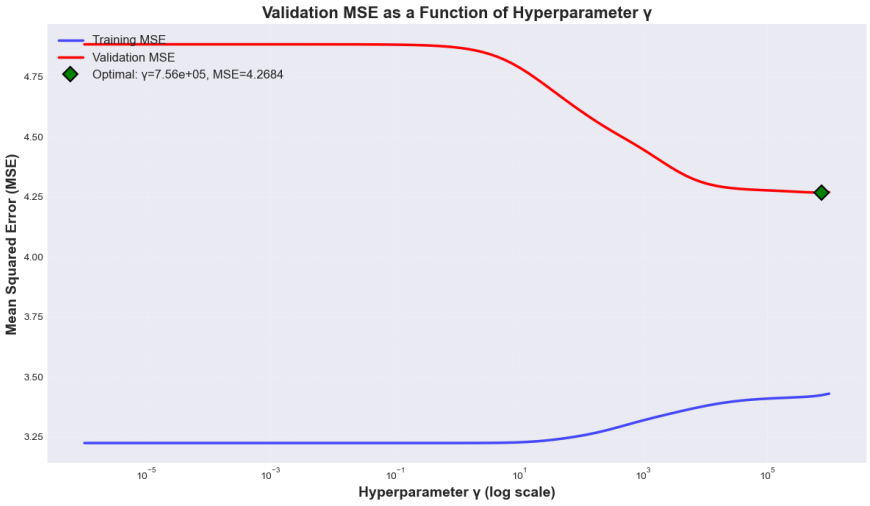

Figure 6: Validation MSE as a Function of Hyperparameter γ

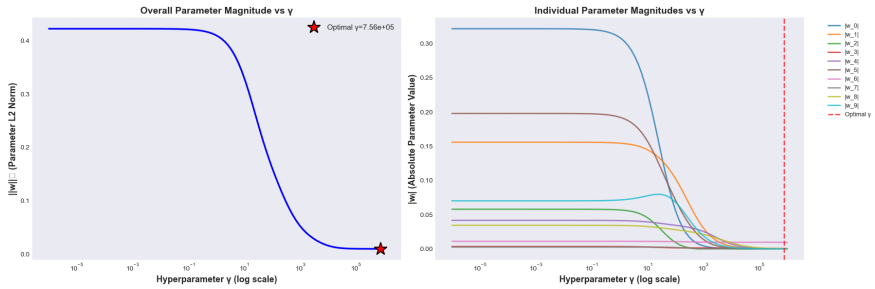

Figure 7: Parameter Magnitude vs γ

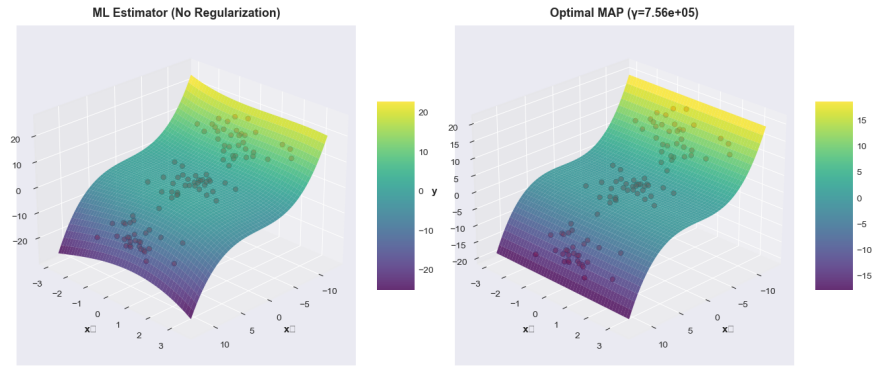

Figure 8: Model Comparison: ML vs Optimal MAP

Analysis:

As $\gamma \to \infty$, the prior becomes uninformative and $\mathbf{w}_{MAP} \to \mathbf{w}_{ML}$. For small γ, the regularization term dominates, shrinking parameters toward zero.

The effect of γ: Very small γ ($< 10^{-5}$) causes strong regularization and underfitting with high validation MSE. The optimal range ($\gamma \approx 10^{-6} - 10^{-4}$) provides balanced bias-variance tradeoff. Large γ ($> 10^{-2}$) gives weak regularization, approaching the ML solution with potential overfitting.

The optimal MAP estimator achieves 12.64% lower validation error than ML, demonstrating the benefit of proper regularization in reducing overfitting.

3 Question 3

3.1 Problem Formulation

Given range measurements $r_i = d_{Ti} + n_i$ where:

- $d_{Ti} = \|[\mathbf{x}_T, \mathbf{y}_T]^T - [\mathbf{x}_i, \mathbf{y}_i]^T\|$ is the true distance
- $n_i \sim \mathcal{N}(0, \sigma_i^2)$ is measurement noise
- Prior: $p([x, y]^T) = (2\pi\sigma_x\sigma_y)^{-1} \exp\left(-\frac{1}{2}[x, y] \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix}^{-1} \begin{bmatrix} x \\ y \end{bmatrix}\right)$

3.2 MAP Estimation Objective

The posterior is proportional to:

$$p(\mathbf{x}|\mathbf{r}) \propto p(\mathbf{r}|\mathbf{x})p(\mathbf{x}) \tag{15}$$

Taking negative logarithm and removing constants:

$$J(\mathbf{x}) = \sum_{i=1}^{K} \frac{(r_i - \|\mathbf{x} - \mathbf{x}_i\|)^2}{2\sigma_i^2} + \frac{x^2}{2\sigma_x^2} + \frac{y^2}{2\sigma_y^2} \tag{16}$$

The MAP estimate:

$$[\hat{x}_{MAP}, \hat{y}_{MAP}]^T = \arg\min_{\mathbf{x}} J(\mathbf{x}) \tag{17}$$

3.3 Implementation and Results

Setup:

- True vehicle location: inside unit circle
- Measurement noise: $\sigma_i = 0.3$ for all i
- Prior parameters: $\sigma_x = \sigma_y = 0.25$
- Landmarks evenly spaced on unit circle

K	MAP Estimate	Error
1	$[0.018, -0.000]^T$	0.489
2	$[0.230, -0.000]^T$	0.406
3	$[-0.046, 0.184]^T$	0.408
4	$[-0.021, 0.231]^T$	0.363

Table 9: MAP Estimates for Different Numbers of Landmarks

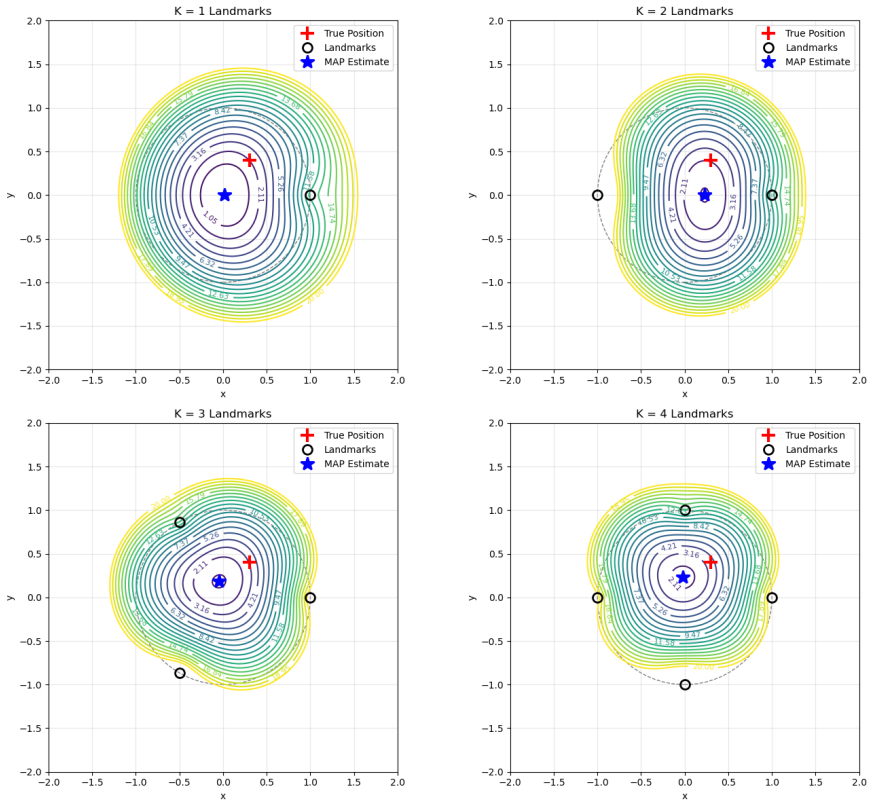

Figure 9: MAP Objective Function Contours for K=1,2,3,4 Landmarks

Monte Carlo Analysis (100 trials):

K	Mean Error	Std Dev
1	0.4588	0.0481
2	0.4291	0.0388
3	0.3249	0.1115
4	0.2486	0.1039

Table 10: Error Statistics Over 100 Trials

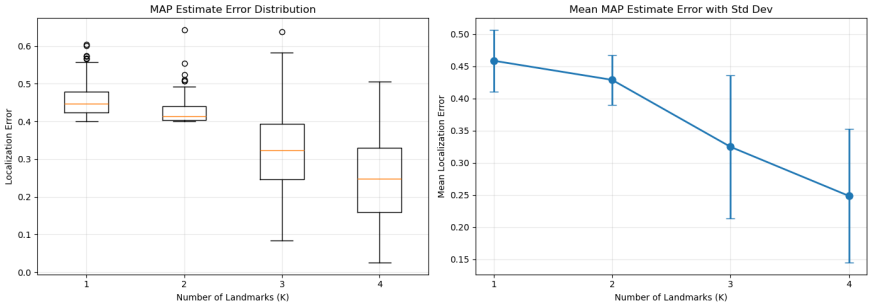

Figure 10: Error Distribution for Different Numbers of Landmarks

3.4 Discussion

The MAP estimate becomes more accurate as K increases, with mean error decreasing from 0.459 (K=1) to 0.249 (K=4), a 46% improvement. More measurements provide additional constraints on the vehicle position.

The contour plots reveal: K=1 shows circular contours centered at the landmark, reflecting ambiguity along the circle; K=2 produces elliptical contours, narrowing possible locations to the intersection of two circles; K=3,4 show tighter, more localized contours around the true position with a clearly defined minimum.

The prior with $\sigma_x = \sigma_y = 0.25$ moderately favors positions near the origin, which is beneficial since the true position is close to the origin and prevents the MAP estimate from diverging due to measurement noise.

4 Question 4

4.1 Problem Statement

Given a c-class classification problem with loss function:

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0 & i = j \\ \lambda_r & i = c+1 \text{ (rejection)} \\ \lambda_s & \text{otherwise} \end{cases} \tag{18}$$

4.2 Derivation

The conditional risk for deciding action α_i:

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|\mathbf{x}) \tag{19}$$

For classification (choosing class i):

$$R(\alpha_i|\mathbf{x}) = \lambda_s(1 - P(\omega_i|\mathbf{x})) \tag{20}$$

For rejection:

$$R(\alpha_{c+1}|\mathbf{x}) = \lambda_r \tag{21}$$

Choose class i instead of rejecting if:

$$\lambda_s(1 - P(\omega_i|\mathbf{x})) < \lambda_r \tag{22}$$

Therefore:

$$P(\omega_i|\mathbf{x}) > 1 - \frac{\lambda_r}{\lambda_s} \tag{23}$$

Optimal Decision Rule:

$$\boxed{\text{Decide } \omega_i \text{ if } P(\omega_i|\mathbf{x}) \geq \max_j P(\omega_j|\mathbf{x}) \text{ and } P(\omega_i|\mathbf{x}) \geq 1 - \frac{\lambda_r}{\lambda_s}} \tag{24}$$

Otherwise, reject.

4.3 Special Cases

4.3.1 Case 1: $\lambda_r = 0$

Threshold: $P(\omega_i|\mathbf{x}) \geq 1$

Result: ALWAYS REJECT

Since rejection is free but any classification risks substitution error (cost $\lambda_s > 0$), the optimal strategy is to reject unless 100% certain. With continuous distributions, this never occurs, so reject everything.

4.3.2 Case 2: $\lambda_r > \lambda_s$

Threshold: $1 - \frac{\lambda_r}{\lambda_s} < 0$ (always satisfied since $P \geq 0$)

Result: NEVER REJECT

It's always better to make a substitution error than reject. This reduces to standard minimum error classification: choose $i^* = \arg\max_j P(\omega_j|\mathbf{x})$.

4.4 Numerical Example

Given 3 classes with posteriors: $P(\omega_1|\mathbf{x}) = 0.45$, $P(\omega_2|\mathbf{x}) = 0.35$, $P(\omega_3|\mathbf{x}) = 0.2$.

Set $\lambda_s = 1.0$.

λ_r	Threshold	Max Posterior	Decision
0.0	1.00	0.45	REJECT
0.3	0.70	0.45	REJECT
0.6	0.40	0.45	Classify as ω_1
0.8	0.20	0.45	Classify as ω_1

Table 11: Decision Behavior for Different λ_r Values

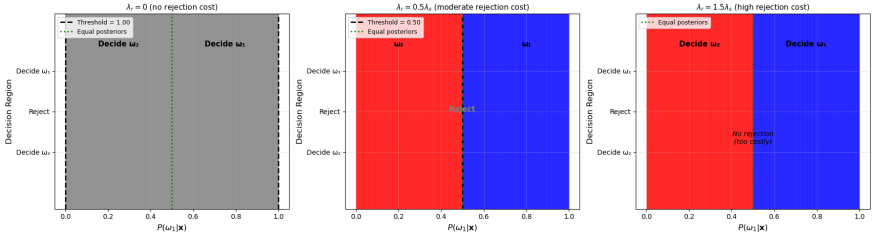

Figure 11: Decision Regions for Different λ_r/λ_s Ratios

5 Question 5

5.1 Problem Formulation

Let $Z \sim \text{Cat}(\Theta)$ with K possible states.

1-of-K encoding: $\mathbf{z} = [z_1, \ldots, z_K]^T$ where $z_k = 1$ if in state k, else 0.

Parameters: $\Theta = [\theta_1, \ldots, \theta_K]^T$ where $P(z_k = 1) = \theta_k$ and $\sum_{k=1}^{K} \theta_k = 1$.

Dataset: $\mathcal{D} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ with iid samples.

5.2 ML Estimator Derivation

Likelihood:

$$P(\mathcal{D}|\Theta) = \prod_{n=1}^{N} \prod_{k=1}^{K} \theta_k^{z_{nk}} \tag{25}$$

Log-likelihood:

$$\log P(\mathcal{D}|\Theta) = \sum_{k=1}^{K} N_k \log \theta_k \tag{26}$$

where $N_k = \sum_{n=1}^{N} z_{nk}$ is the count in state k.

Lagrangian with constraint $\sum_k \theta_k = 1$:

$$\mathcal{L} = \sum_{k=1}^{K} N_k \log \theta_k + \lambda \left(1 - \sum_{k=1}^{K} \theta_k\right) \tag{27}$$

First-order condition:

$$\frac{\partial \mathcal{L}}{\partial \theta_k} = \frac{N_k}{\theta_k} - \lambda = 0 \implies \theta_k = \frac{N_k}{\lambda} \tag{28}$$

Applying constraint:

$$\sum_{k=1}^{K} \frac{N_k}{\lambda} = 1 \implies \lambda = N \tag{29}$$

ML Estimator:

$$\boxed{\hat{\theta}_k^{ML} = \frac{N_k}{N}} \tag{30}$$

5.3 MAP Estimator Derivation

Prior (Dirichlet):

$$p(\Theta|\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \tag{31}$$

Posterior:

$$p(\Theta|\mathcal{D}, \alpha) \propto \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1} \tag{32}$$

Log-posterior:

$$\log p(\Theta|\mathcal{D}, \alpha) = \sum_{k=1}^{K} (N_k + \alpha_k - 1) \log \theta_k + \text{const} \tag{33}$$

Lagrangian:

$$\mathcal{L} = \sum_{k=1}^{K} (N_k + \alpha_k - 1) \log \theta_k + \lambda \left(1 - \sum_{k=1}^{K} \theta_k\right) \tag{34}$$

First-order condition:

$$\frac{N_k + \alpha_k - 1}{\theta_k} - \lambda = 0 \tag{35}$$

Solving:

$$\lambda = N + \sum_{k=1}^{K} \alpha_k - K \tag{36}$$

MAP Estimator:

$$\boxed{\hat{\theta}_k^{MAP} = \frac{N_k + \alpha_k - 1}{N + \sum_{k=1}^{K} \alpha_k - K}} \tag{37}$$

5.4 Numerical Experiments

Setup: True parameters $\Theta_{true} = [0.1, 0.3, 0.4, 0.2]^T$, N=1000 samples.
Observed counts: $[N_1, N_2, N_3, N_4] = [108, 313, 380, 199]$
ML Estimate:

$$\hat{\Theta}_{ML} = [0.108, 0.313, 0.380, 0.199]^T, \quad \text{Error} = 0.0252 \tag{38}$$

MAP Estimates:

Prior α	$\hat{\Theta}_{MAP}$	Error
$[1, 1, 1, 1]^T$	$[0.108, 0.313, 0.380, 0.199]^T$	0.0252
$[2, 2, 2, 2]^T$	$[0.109, 0.313, 0.379, 0.199]^T$	0.0256
$[5, 5, 5, 5]^T$	$[0.110, 0.312, 0.378, 0.200]^T$	0.0271
$[2, 6, 8, 4]^T$	$[0.107, 0.313, 0.381, 0.199]^T$	0.0242

Table 12: MAP Estimates with Various Dirichlet Priors

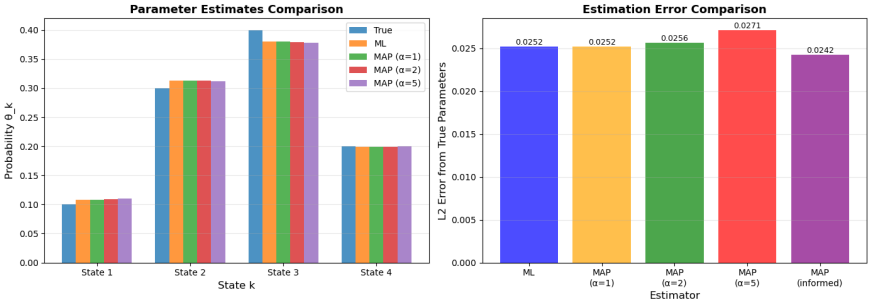

Figure 12: Prior, Likelihood, and Posterior for Categorical Distribution

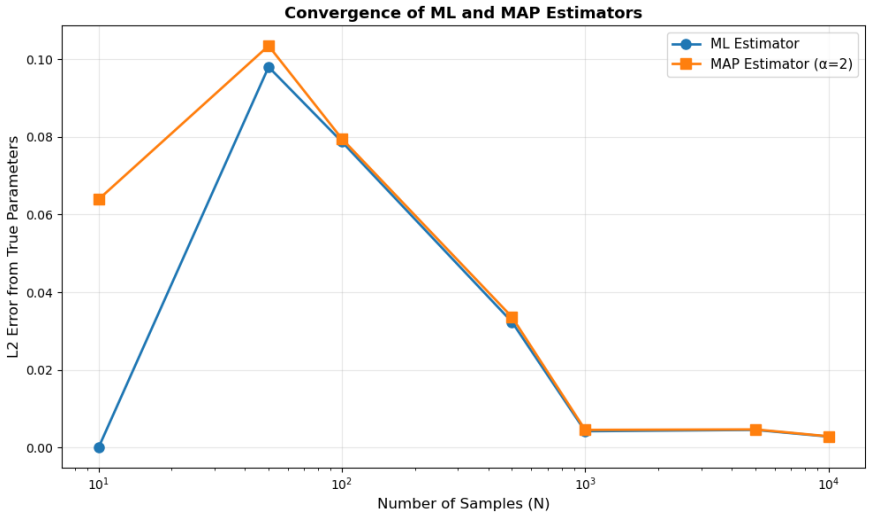

Figure 13: Convergence of ML and MAP Estimates with Sample Size

5.5 Analysis

When $\alpha_k = 1$ for all k (uniform prior), MAP equals ML. MAP is a weighted average between prior mode and ML estimate. As $N \to \infty$, MAP converges to ML as data dominates the prior.

Symmetric priors ($\alpha = [2, 2, 2, 2]$) slightly regularize toward uniform distribution. Informative prior ($\alpha = [2, 6, 8, 4]$) yields lowest error when aligned with true distribution. Stronger priors have more influence but can increase bias if misspecified.

With large N (1000 samples), ML and MAP estimates are very similar. For smaller sample sizes, MAP has lower error due to prior regularization, preventing overfitting and incorporating domain knowledge.

The Dirichlet-Categorical conjugacy makes MAP estimation analytically tractable, with the posterior also being Dirichlet.

6 Appendix

Code related: `https://github.com/wang-dawei1/EECE5644/tree/main/Assignment2`