Report for Project 1

Md Sultan Mahmud

E-mail: mqm7099@psu.edu

1. Derivations

Given N data points and an M-degree polynomial as our curve fitting model, let

$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_M \end{bmatrix}, \quad \mathbf{t} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix}, \quad \text{and} \quad \mathbf{X} = \begin{bmatrix} x_1^0 & x_1^1 & \cdots & x_1^M \\ x_2^0 & x_2^1 & \cdots & x_2^M \\ \vdots & \vdots & \ddots & \vdots \\ x_N^0 & x_N^1 & \cdots & x_N^M \end{bmatrix},$$

where $\mathbf{w}$ is an $(M+1) \times 1$ weight vector, $\mathbf{t}$ is an $N \times 1$ target vector, and $\mathbf{X}$ is an $N \times (M+1)$ input matrix.

1.1. ML estimator from a probabilistic perspective

The log likelihood function is (refer to Bishop [1] Equation 1.62) –

$$\ln p(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{y(\mathbf{x}_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi). \tag{1}$$

Let's determine the maximum likelihood solution for the polynomial coefficients, which will be denoted by $\mathbf{w}_{ML}$.

Remark:

(i) In **1**: we can omit the last two terms on the right-hand side as they do not depend on $\mathbf{w}$.

(ii) Also, we note that scaling the log likelihood by a positive constant coefficient does not alter the location of the maximum with respect to $\mathbf{w}$, and so we can replace the coefficient $\beta/2$ with $1/2$.

(iii) Finally, instead of maximizing the log likelihood, we can equivalently minimize the negative log likelihood. We therefore see that maximizing likelihood is equivalent, so far as determining $\mathbf{w}$ is concerned, to minimizing the *sum-of-squares error function* $E(\mathbf{w})$ defined by Bishop [1] Equation 1.2.

$$E(\mathbf{w}) = \frac{1}{2}(\mathbf{X}\mathbf{w} - \mathbf{t})^\top(\mathbf{X}\mathbf{w} - \mathbf{t}). \tag{2}$$

$$= \frac{1}{2}(\mathbf{w}^\top\mathbf{X}^\top - \mathbf{t}^\top)(\mathbf{X}\mathbf{w} - \mathbf{t}). \tag{3}$$

$$= \frac{1}{2}(\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} - \mathbf{w}^\top\mathbf{X}^\top\mathbf{t} - \mathbf{t}^\top\mathbf{X}\mathbf{w} - \mathbf{t}^\top\mathbf{t}). \tag{4}$$

$$= \frac{1}{2}(\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} - 2\mathbf{w}^\top\mathbf{X}^\top\mathbf{t} - \mathbf{t}^\top\mathbf{t}). \tag{5}$$

Remark:

(i) From **2** to **3**: use the transpose rule of matrix.

(ii) From **3** to **4**: use the distribution rule of matrix.

(iii) From **4** to **5**: notice that $\mathbf{w}^\top\mathbf{X}^\top\mathbf{t}$ and $\mathbf{t}^\top\mathbf{X}\mathbf{w}$ are both scalars. And the **transpose of a scalar is itself**. Use this transpose rule on $\mathbf{t}^\top\mathbf{X}\mathbf{w}$, we have

$$\mathbf{t}^\top\mathbf{X}\mathbf{w} = \mathbf{w}^\top\mathbf{X}^\top\mathbf{t},$$

and combine like terms we obtain

$$-2\mathbf{w}^\top\mathbf{X}^\top\mathbf{t}.$$

Taking the derivative of $\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}}$:

$$\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} = \frac{1}{2}\left(\frac{\partial}{\partial \mathbf{w}}\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} - 2\frac{\partial}{\partial \mathbf{w}}\mathbf{w}^\top\mathbf{X}^\top\mathbf{t} - \frac{\partial}{\partial \mathbf{w}}\mathbf{t}^\top\mathbf{t}\right). \tag{6}$$

$$= \frac{1}{2}\left(2\mathbf{X}^\top\mathbf{X}\mathbf{w} - 2\mathbf{X}^\top\mathbf{t}\right). \tag{7}$$

$$= \mathbf{X}^\top\mathbf{X}\mathbf{w} - \mathbf{X}^\top\mathbf{t}. \tag{8}$$

Remark:

(i) From **6** to **7** (on the first term: $\frac{\partial}{\partial \mathbf{w}}\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w}$): notice that $\mathbf{X}^\top\mathbf{X}$ is symmetric, and use the rule if $\mathbf{A}$ is symmetric, $\frac{\partial(\mathbf{x}^\top\mathbf{A}\mathbf{x})}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$.

(ii) From **6** to **7** (on the second term: $\frac{\partial}{\partial \mathbf{w}}\mathbf{w}^\top\mathbf{X}^\top\mathbf{t}$): use the rule $\frac{\partial \mathbf{x}^\top\mathbf{a}}{\partial \mathbf{x}} = \mathbf{a}$.

(iii) From **6** to **7** (on the third term: $\frac{\partial}{\partial \mathbf{w}}\mathbf{t}^\top\mathbf{t}$): The $\mathbf{t}^\top\mathbf{t}$ drops out w.r.t. $\mathbf{w}$.

Therefore, the analytical solution of optimal $\mathbf{w}^*$ can be obtained by making:

$$\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}^*} = 0$$

Then we solve for $\mathbf{w}^*$.

$$\mathbf{X}^\top\mathbf{X}\mathbf{w}^* - \mathbf{X}^\top\mathbf{t} = 0 \tag{9}$$

$$\mathbf{w}^* = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{t} \tag{10}$$

Therefore, the maximum likelihood solution for the polynomial coefficients, $\mathbf{w}_{ML}$ is :

$$(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{t}$$

1.2. MAP estimator of the Bayesian approach

The log of the posterior distribution is given by the sum of the log likelihood and the log of the prior and, as a function of $\mathbf{w}$, takes the form (refer to Bishop [1] Equation 3.55):

$$\ln p(\mathbf{w} \mid \mathbf{t}) = -\frac{\beta}{2}\sum_{n=1}^{N}\left\{t_n - \mathbf{w}^\top\phi(\mathbf{x}_n)\right\}^2 - \frac{\alpha}{2}\mathbf{w}^\top\mathbf{w} + \text{const.} \tag{11}$$

Maximization of this posterior distribution with respect to $\mathbf{w}$ is therefore equivalent to the minimization of the sum-of-squares error function with the addition of a quadratic regularization term, corresponding to Bishop [1] Equation 1.67.

$$E(\mathbf{w}) = \frac{\beta}{2}\sum_{n=1}^{N}\left\{y(\mathbf{x}_n, \mathbf{w}) - t_n\right\}^2 + \frac{\alpha}{2}\mathbf{w}^\top\mathbf{w}. \tag{12}$$

As done in the previous section, for computational ease, we rewrite **12** in terms of matrix notation as:

$$\frac{\partial E}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}}\left((\mathbf{X}\cdot\mathbf{w} - \mathbf{t})^2 + \frac{\alpha}{2}\mathbf{w}^\top\mathbf{w}\right) = 0 \tag{13}$$

It can also be rewritten as:

$$\frac{\partial}{\partial \mathbf{w}}\left((\mathbf{X}\mathbf{w} - \mathbf{t})^\top(\mathbf{X}\mathbf{w} - \mathbf{t}) + \frac{\alpha}{2}\mathbf{w}^\top\mathbf{w}\right) = 0 \tag{14}$$

Here, $\mathbf{t}$ simply represents the ground truth value for the regression. Given the variables, **14** can be expanded using matrix calculation as follows:

$$\begin{aligned}&\frac{\partial}{\partial \mathbf{w}}\left(\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} - \mathbf{w}^\top\mathbf{X}^\top\mathbf{T} - \mathbf{T}^\top\mathbf{X}\mathbf{w} + \mathbf{T}^\top\mathbf{T} + \frac{\alpha}{2}\mathbf{w}^\top\mathbf{w}\right)\\ &= \frac{\partial}{\partial \mathbf{w}}\left(\mathbf{w}^\top\left(\mathbf{X}^\top\mathbf{X} - \mathbf{X}^\top\mathbf{T} + \frac{\alpha}{\beta}\mathbf{w}\right) - \mathbf{T}^\top(\mathbf{X}\mathbf{w} - \mathbf{T})\right) = 0\end{aligned} \tag{15}$$

The regression model aims to minimize the error (difference) between the predicted values ($\mathbf{X}\mathbf{w}$) and the actual values ($\mathbf{t}$). By solving for $\mathbf{X}\mathbf{w} - \mathbf{t} = 0$, we are finding the weights ($\mathbf{w}$) that minimize this residual error.

Holding $(\mathbf{X}\mathbf{w} - \mathbf{t}) = 0$, we get:

$$\left(\mathbf{X}^\top\mathbf{X}\mathbf{w} - \mathbf{X}^\top\mathbf{t} + \frac{\alpha}{\beta}\mathbf{w}\right) = 0 \tag{16}$$

Then we solve for $\mathbf{w}$.

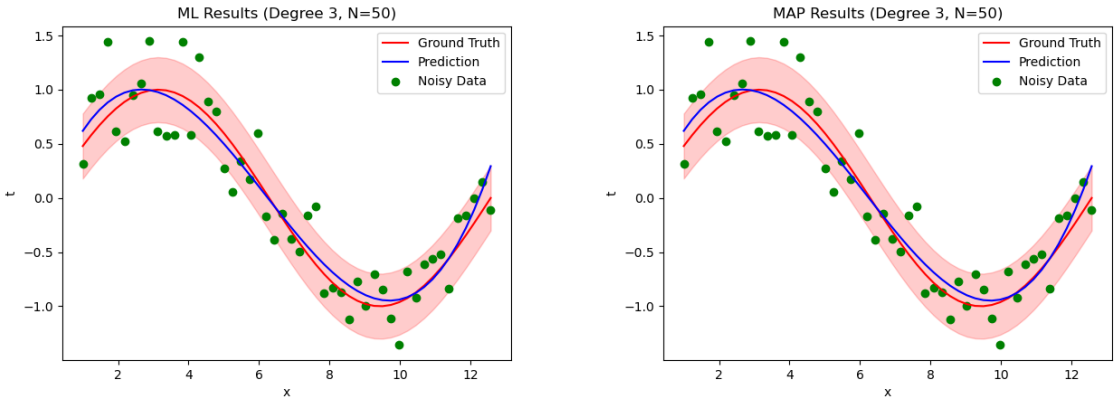
$$\mathbf{w} = \left(\beta\mathbf{X}^\top\mathbf{X} + \alpha\mathbf{I}\right)^{-1}\beta\mathbf{X}^\top\mathbf{t}. \tag{17}$$

Therefore, the maximum a posteriori solution for the polynomial coefficients, $\mathbf{w}_{MAP}$ is :

$$\left(\beta\mathbf{X}^\top\mathbf{X} + \alpha\mathbf{I}\right)^{-1}\beta\mathbf{X}^\top\mathbf{t}.$$

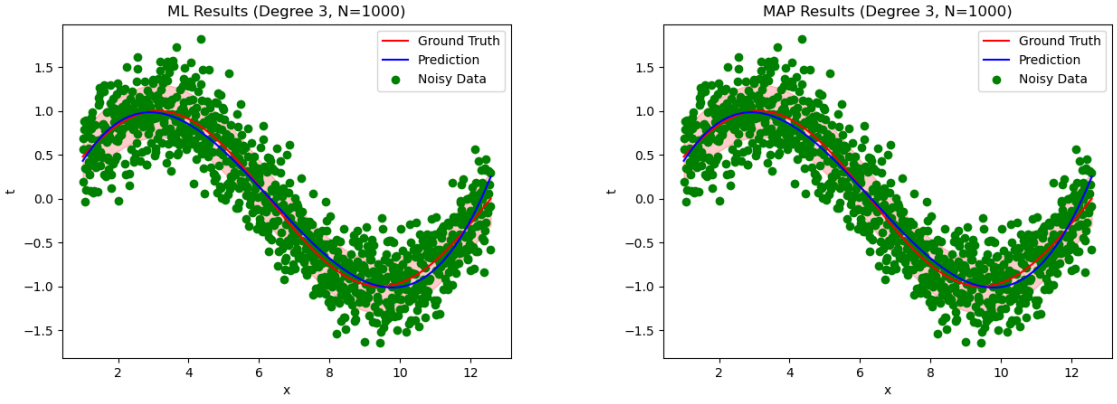
2. Results for the estimated regression models

2.1. Analysis of Polynomial Degrees



(a) ML (b) MAP

Figure 1: ML and MAP model predictions for a polynomial degree of 3 with N=50 data points. This figure demonstrates the simplicity and smoothness of the low-degree polynomial fit. Both models produce similar predictions due to the low complexity, avoiding overfitting but resulting in underfitting for noisy datasets

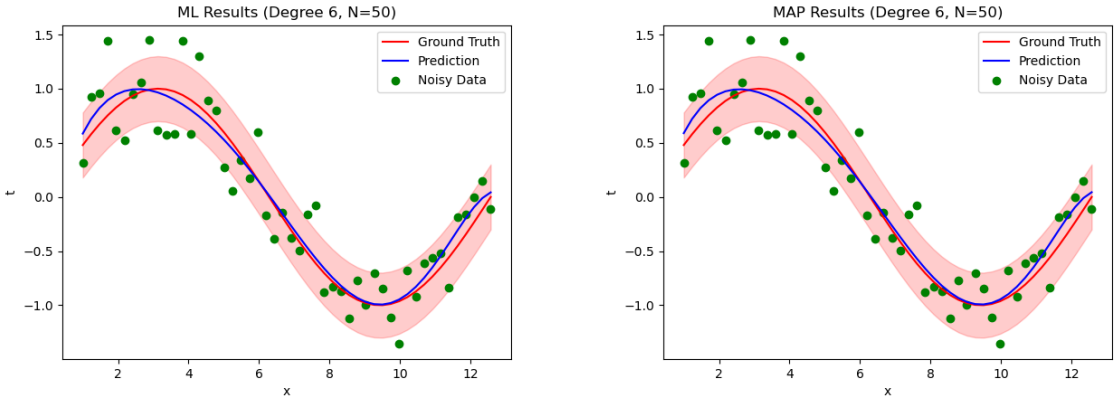


(a) ML (b) MAP

Figure 2: ML and MAP model predictions for a polynomial degree of 3 with N=1000 data points. The larger dataset provides more information, enabling both models to follow the ground truth more closely while maintaining simplicity and avoiding overfitting

2.1.1. Degree $M = 3$:

- A low-degree polynomial provides a simple and smooth fit.
- Both ML and MAP models produce similar predictions, as the low complexity avoids overfitting.
- The model struggles to capture finer details of the data, leading to underfitting, especially for noisy datasets.



(a) ML (b) MAP

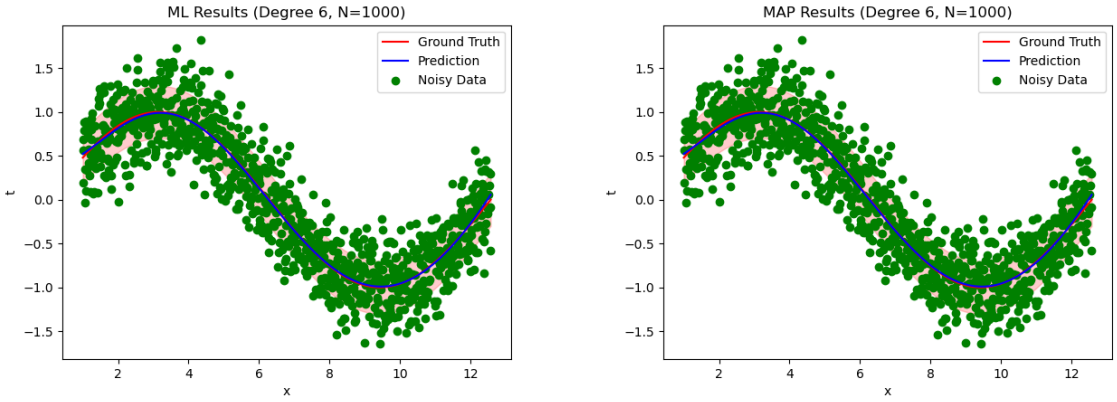
Figure 3: ML and MAP model predictions for a polynomial degree of 6 with N=50 data points. The higher degree captures more data structure, balancing between fitting the sine function and avoiding excessive noise sensitivity. MAP's regularization ensures smoother predictions compared to ML

2.1.2. Degree $M = 6$:

- A moderate-degree polynomial captures more of the data's structure compared to $M = 3$.
- The models balance between fitting the underlying sine function and avoiding excessive sensitivity to noise.
- Predictions improve in terms of following the ground truth curve more closely without overfitting.

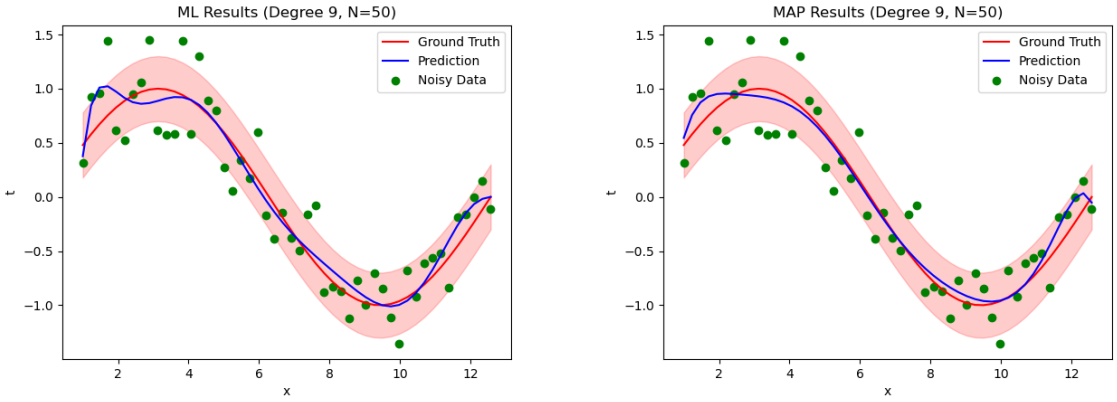
2.1.3. Degree $M = 9$:

- A high-degree polynomial allows for greater flexibility, enabling the model to capture finer details in the data.



(a) ML (b) MAP

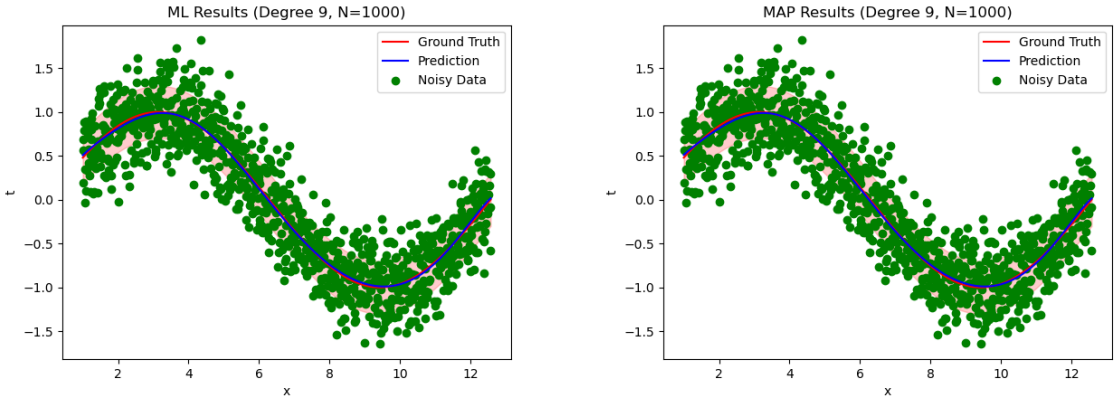
Figure 4: ML and MAP model predictions for a polynomial degree of 6 with N=1000 data points. With a larger dataset, both models align closely with the ground truth curve, effectively handling the higher polynomial complexity without overfitting



(a) ML (b) MAP

Figure 5: ML and MAP model predictions for a polynomial degree of 9 with N=50 data points. The high-degree polynomial offers greater flexibility, allowing ML to overfit noise in the small dataset, whereas MAP's regularization mitigates overfitting, producing smoother predictions

- For smaller datasets ($N = 50$), ML tends to overfit by following the noise in the data, while MAP's regularization mitigates this to some extent.
- For larger datasets ($N = 1000$), the abundance of data allows both ML and MAP to handle higher degrees effectively, aligning closely with the ground truth without significant overfitting.



(a) ML (b) MAP

Figure 6: ML and MAP model predictions for a polynomial degree of 9 with N=1000 data points. The large dataset supports the higher degree, enabling both models to closely approximate the ground truth without significant overfitting

2.1.4. Key Observations:

- Increasing M provides greater flexibility but risks overfitting, especially for small datasets.
- MAP's regularization helps control overfitting at higher degrees ($M = 9$), particularly when the dataset is small.
- For larger datasets, both ML and MAP perform well across all degrees as the noise influence diminishes.

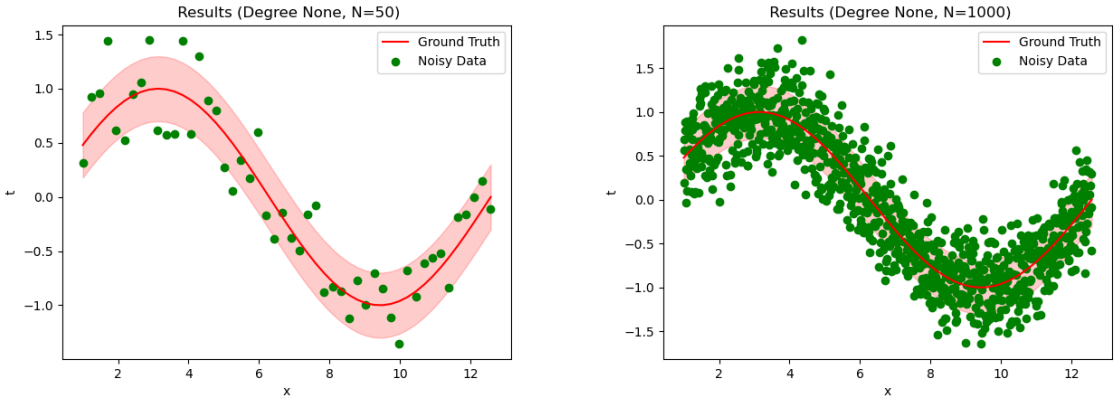
2.2. Clean vs. Noisy Data

2.2.1. Clean Data (Ground Truth):

- The clean data represents the sine wave ($y = \sin(0.5x)$) without any noise.
- Both ML and MAP models aim to approximate this curve based on the observed data.
- Clean data provides the baseline for evaluating the effectiveness of the models in capturing the underlying trend.

2.2.2. Noisy Data:

- **Small Dataset ($N = 50$):**
  - The limited number of data points amplifies the effect of noise, making it harder for both models to distinguish between the true trend and noise.



(a) N=50 (b) N=1000

Figure 7: Comparison of clean and noisy datasets for two dataset sizes: N=50 and N=1000. The sparse dataset (N=50) highlights noise influence, making it harder for models to generalize effectively, whereas the dense dataset (N=1000) reduces noise impact, leading to smoother predictions

  - ML tends to overfit the noise, especially at higher polynomial degrees ($M = 9$), while MAP's regularization prevents overfitting by enforcing smoother predictions.

- **Large Dataset ($N = 1000$):**
  - The higher number of data points reduces the influence of noise as the dataset more accurately samples the sine wave.
  - Both models produce smoother predictions that align closely with the clean data, as the noise averages out due to the Central Limit Theorem.

2.2.3. Key Observations:

- **Small datasets are more sensitive to noise**, making it difficult for the models to generalize effectively. MAP's regularization makes it more robust in these scenarios.
- **Larger datasets mitigate noise**, allowing both models to better approximate the clean data.
- The presence of noise challenges both models, but MAP's ability to regularize ensures more reliable predictions, especially with sparse data.

2.3. Size of Different Dataset Points (N)

2.3.1. Small Dataset ($N = 50$):

- The limited number of data points makes the dataset sparse, increasing the impact of noise on predictions.
- The models have fewer data points to learn from, which can result in overfitting, especially for higher polynomial degrees ($M = 9$).
- MAP performs better in controlling overfitting due to its regularization, producing smoother predictions compared to ML.
- Predictions tend to be less reliable and more variable due to the lack of sufficient data coverage across the input range.

2.3.2. Large Dataset ($N = 1000$):

- The abundance of data points provides a dense sampling of the input space, reducing the influence of noise.
- Both ML and MAP models perform effectively, producing smoother predictions that align closely with the ground truth.
- Regularization in MAP becomes less critical as the noise influence diminishes due to the larger sample size.
- Overfitting is less of a concern for higher polynomial degrees ($M = 9$), as the large dataset provides enough information to support the model's complexity.

2.3.3. Key Observations:

- **Small datasets ($N = 50$)** are prone to noise and overfitting, making MAP's regularization advantageous.
- **Large datasets ($N = 1000$)** provide sufficient information for both models to generalize effectively, reducing the impact of noise and overfitting.
- Increasing N consistently improves the stability and reliability of predictions for both models, with MAP and ML converging to similar performance as the dataset size grows.

2.4. Extra Credit)

2.4.1. Key Observations

- **Polynomial Degree (M):**
  - Increasing M reduces RMSE by capturing more complexity in the data.
  - At $M = 9$, ML achieves slightly lower RMSE due to overfitting, while MAP's regularization penalizes this overfitting.
- **Dataset Size (N):**

Degree (M)	**ML RMSE** ($N = 50$)	**MAP RMSE** ($N = 50$)	**ML RMSE** ($N = 1000$)	**MAP RMSE** ($N = 1000$)
3	0.2825	0.2825	0.3090	0.3090
6	0.2758	0.2758	0.3042	0.3042
9	0.2650	0.2702	0.3041	0.3041

Table 1: RMSE values for ML and MAP models across different polynomial degrees (M=3, 6, 9) and dataset sizes (N=50 and N=1000). The table highlights the trade-off between flexibility and regularization. MAP exhibits robustness in small datasets by mitigating overfitting, while both models converge in performance for larger datasets

  - Smaller datasets ($N = 50$) exhibit higher variability due to noise, making MAP more robust at higher degrees.
  - Larger datasets ($N = 1000$) stabilize RMSE, and both models perform nearly identically as noise influence diminishes.

2.4.2. Overall Trends

- For small datasets, MAP is more reliable due to regularization, while ML may overfit at higher degrees.
- For larger datasets, both models converge in performance as overfitting becomes less significant.

2.5. Verification of Central Limit Theorem (CLT)

- **CLT Principle:** As the dataset size (N) increases, the influence of noise diminishes, and the predictions become smoother and align more closely with the ground truth. The sampling distribution of noisy data averages approaches a normal distribution.
- **Observations:**
  - **For $N = 50$:** Predictions are more variable due to sparse data and the higher influence of noise. Both ML and MAP struggle to generalize effectively, especially at higher degrees.
  - **For $N = 1000$:** The larger dataset provides dense sampling, reducing noise influence and leading to smoother predictions. ML and MAP perform nearly identically as the noise averages out.

3. Summary

This project explored polynomial regression using Maximum Likelihood (ML) and Maximum A Posteriori (MAP) models, analyzing their performance across varying polynomial degrees ($M = 3, 6, 9$) and dataset sizes ($N = 50, 1000$). The models were

evaluated using Root Mean Squared Error (RMSE), highlighting key trade-offs between ML's flexibility and MAP's regularization.

For smaller datasets ($N = 50$), MAP demonstrated robustness by mitigating overfitting at higher degrees, while ML achieved slightly lower RMSE by fitting noise. For larger datasets ($N = 1000$), the noise influence diminished, and both models converged in performance. The Central Limit Theorem was validated as larger datasets resulted in smoother predictions and reduced noise variability.

This study underscores the importance of regularization in handling noise and dataset size in ensuring model generalization.

References

[1] Bishop C M and Nasrabadi N M 2006 *Pattern recognition and machine learning* vol 4 (Springer)