Project #1 - Linear Regression

Hayden Moore (hmm5731@psu.edu)
Pattern Recognition and Machine Learning

1. **Derived equations for regression using Maximum Likelihood Estimation (MLE) and Maximum A Posteriori (MAP)**

   MLE given by: $\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

   (a) Model assumption: $\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$

   (b) $\mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$, meaning we can break this down into two parts for the current observed data y: The true underlying relationship we are trying to find ($\mathbf{X}\mathbf{w}$) and noise

   (c) The likelihood function: $p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) = \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I})$, representing a Gaussian distribution with the observed data $\mathbf{y}$ modeled as being normally distributed around the linear prediction $\mathbf{X}\mathbf{w}$, with some variance

   (d) Log likelihood: $\ln p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}||\mathbf{y} - \mathbf{X}\mathbf{w}||^2$

   (e) We take the derivative with respect to $\mathbf{w}$ and set to zero: $\frac{\partial}{\partial \mathbf{w}} \ln p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = 0$ $-\frac{1}{2\sigma^2}\frac{\partial}{\partial \mathbf{w}}(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) = 0$, $\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{y}$

   (f) Finally, we solve for $\mathbf{w}$: $\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

   (g) This solution minimizes the sum of squared errors $||\mathbf{y} - \mathbf{X}\mathbf{w}||^2$

   Similarly, MAP is given by: $\mathbf{w} = (\beta\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I})^{-1}\beta\mathbf{X}^T\mathbf{y}$

   (a) MAP estimation incorporates a prior distribution on the model weights $\mathbf{w}$ to regularize the solution, which helps prevent overfitting.

   (b) The prior on the weights $\mathbf{w}$ is assumed to be Gaussian: $p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \mathbf{0}, \alpha^{-1}\mathbf{I})$, where α controls the regularization strength

   (c) The log posterior distribution, combining the likelihood and prior, is: $\ln p(\mathbf{y}, \mathbf{w} \mid \mathbf{X}) = -\frac{1}{2\sigma^2}||\mathbf{y} - \mathbf{X}\mathbf{w}||^2 - \frac{\alpha}{2}||\mathbf{w}||^2 + \text{constant}$.

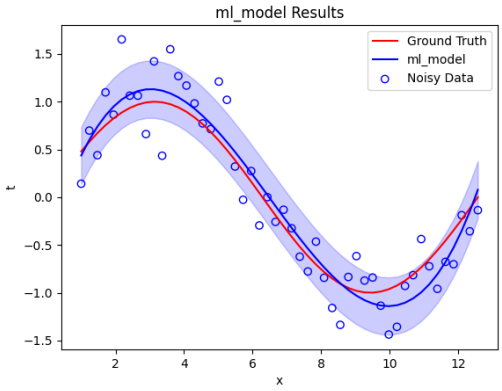
   (d) Taking the derivative with respect to $\mathbf{w}$ and setting it to zero: $\frac{\partial}{\partial \mathbf{w}}\left(-\frac{1}{2\sigma^2}||\mathbf{y} - \mathbf{X}\mathbf{w}||^2 - \frac{\alpha}{2}||\mathbf{w}||^2\right) = 0$,

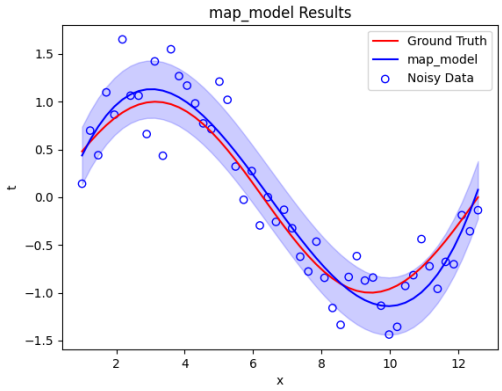
   (e) simplified to: $\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{y} + \alpha\sigma^2\mathbf{w}$.

   (f) Solving for $\mathbf{w}$, we obtain the MAP estimate: $\mathbf{w} = (\beta\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I})^{-1}\beta\mathbf{X}^T\mathbf{y}$.

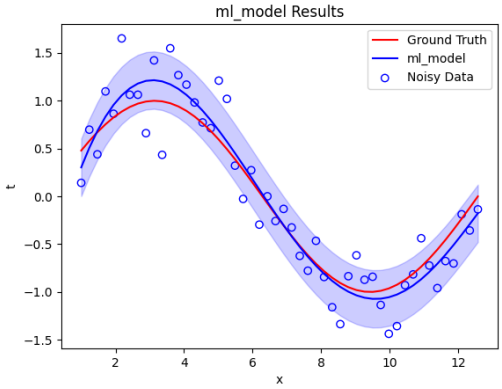
2. **Visualization results for the estimated regression models**

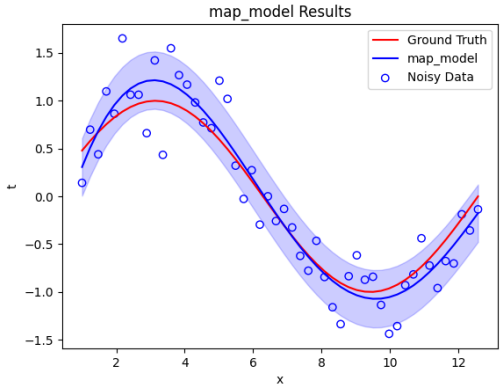
(a) **ML Model (M=3, N=50)** **MAP Model (M=3, N=50)**



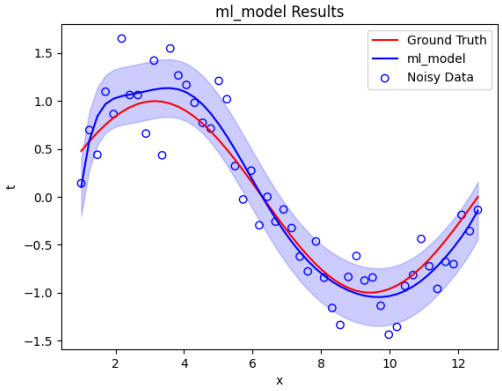


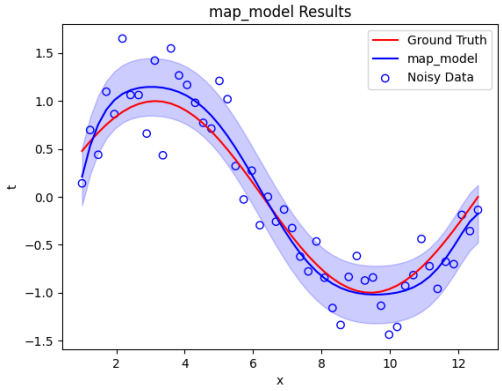
(b) **ML Model (M=6, N=50)** **MAP Model (M=6, N=50)**



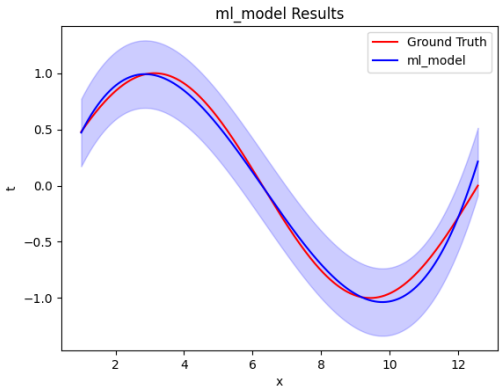


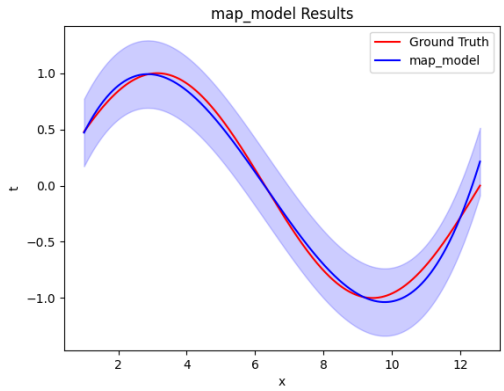
(c) **ML Model (M=9, N=50)** **MAP Model (M=9, N=50)**



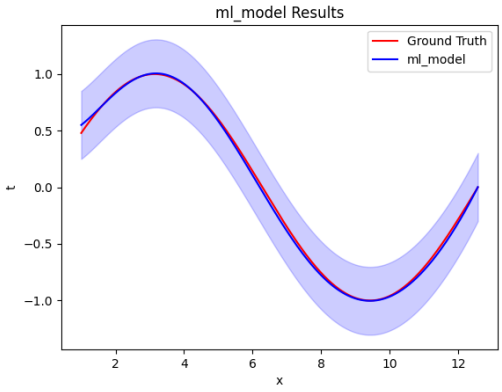


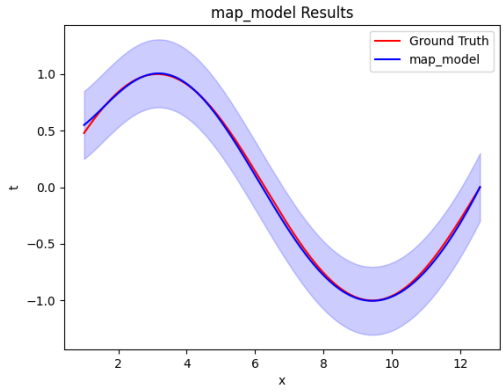
(d) **ML Model (M=3, N=1000)** **MAP Model (M=3, N=1000)**



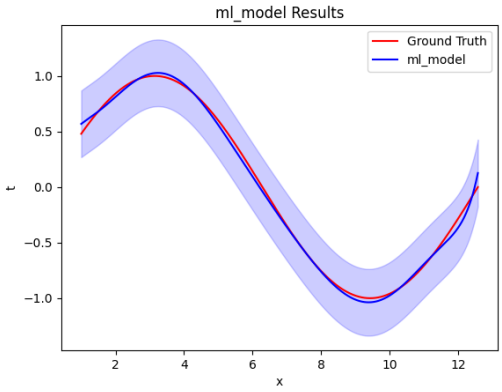


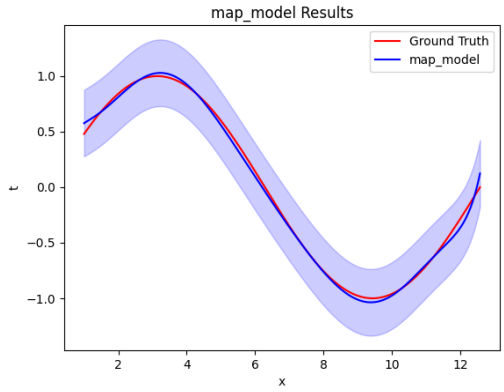
(e) **ML Model (M=6, N=1000)** **MAP Model (M=6, N=1000)**





(f) **ML Model (M=9, N=1000)** **MAP Model (M=9, N=1000)**





3. **A comprehensive summary and comparison of these two methods based on observations of:**

   (a) **Change of degree of polynomial M:**
   Increasing the polynomial degree introduces non-linearity into the model allowing us to better fit the data as we increase this degree (to a point). Here are the differences in performance for MLE and MAP as we increase the degree of the polynomial:

i. With lower polynomial degrees both MLE and MAP perform similarly: Both MLE and MAP aim to fit the data as best as possible, but the low-degree polynomial provides enough simplicity bias in the model that both methods will result in very similar estimates and graphs.

ii. As we increase the polynomial degrees MAP seems to perform slightly better: This makes sense as the regularization terms with MAP help to prevent overfitting.

iii. At polynomial degree nine we begin to see artifacts of overfitting as the curves from the plots start to look less smooth and have minor spikes. We can see this in both MLE and MAP but we have a slightly more smooth graph when looking at MAP.

(b) **Clean vs Noisy data:**
Clean data allows both MLE and MAP to fit more quickly when compared to Noisy data. With clean data we can quickly fit the model at lower polynomial degrees since the data is consistent and expected. Whereas when working with the noisy data MLE and MAP both have to learn an estimated solution that tries to best fit in-between the noise, which takes a higher polynomial degree.

(c) **Size of different data sets points:**
As we increase the data size from N=50 to N=1000 we fit much closer to the ground truth line more quickly (at lower polynomial degrees). This proves that more data even if noisy within some distribution allows both MLE and MAP to fit appropriately and quickly.

4. **Extra Credit Attempt:**
NONE

5. **Verification of Central Limit Theorem:**
The behavior of MLE and MAP in these experiments demonstrates key aspects of the Central Limit Theorem: As the sample size increases from N=50 to N=1000, both MLE and MAP clearly provide better and more stable estimates. Even if the data is noisy, the larger the sample, the closer the model gets to the true underlying distribution at lower and higher polynomial degrees.