Probabilistic View of Linear Regression

Nipun Batra

February 9, 2020

IIT Gandhinagar

Probabilistic View of Linear Regression

- Example function (black solid diagonal line) and its predictive uncertainty at x = 60 (drawn as a Gaussian).

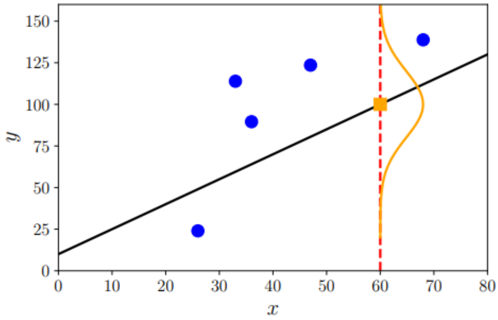


Figure 1: Probabilistic view of Linear Regression. Note that we don't have point estimates any longer.

Probabilistic View of Linear Regression

- In this view, we consider a likelihood function

$$p(y|\boldsymbol{x}) = \mathcal{N}\left(y|f(\boldsymbol{x}), \sigma^2\right)$$

where $\boldsymbol{x} \in \mathbb{R}^D$ and the inputs and $y \in \mathbb{R}$ are the noisy function values, with the functional relationship between $\boldsymbol{x}$ and y given by

$$y = f(\boldsymbol{x}) + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$, is i.i.d. measurement noise with mean 0 and variance σ^2.

Parameter Estimation and MLE

- Suppose we are given a training set $\mathcal{D} := \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_N)$, consisting of N inputs $\boldsymbol{x}_n \in \mathbb{R}^D$ and corresponding targets $y_n \in \mathbb{R}$, $n = 1, 2, 3, \ldots N$. The graphical model for the same under the probabilistic viewpoint is as given below.

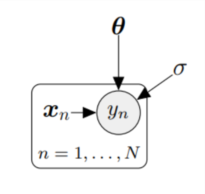


Figure 2: Probabilistic Graphical Model for Linear Regression

In the above PGM, the observed random variables are shaded and the deterministic random variables are without circles.

Parameter Estimation

- Note that y_i and y_j are conditionally independent given their respective inputs $\boldsymbol{x}_i, \boldsymbol{x}_j$ so that the likelihood factorizes according to

$$\begin{aligned} p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta}) &= p\left(y_1, \ldots, y_N | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{\theta}\right) \\ &= \prod_{n=1}^{N} p\left(y_n | \boldsymbol{x}_n, \boldsymbol{\theta}\right) = \prod_{n=1}^{N} \mathcal{N}\left(y_n | \boldsymbol{x}_n^\top \boldsymbol{\theta}, \sigma^2\right) \end{aligned}$$

where $\mathcal{X} := \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ and $\mathcal{Y} := \{y_1, y_2, \ldots, y_n\}$.

Parameter Estimation

- Note that y_i and y_j are conditionally independent given their respective inputs $\boldsymbol{x}_i, \boldsymbol{x}_j$ so that the likelihood factorizes according to

$$\begin{aligned} p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta}) &= p\left(y_1, \ldots, y_N | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{\theta}\right) \\ &= \prod_{n=1}^{N} p\left(y_n | \boldsymbol{x}_n, \boldsymbol{\theta}\right) = \prod_{n=1}^{N} \mathcal{N}\left(y_n | \boldsymbol{x}_n^\top \boldsymbol{\theta}, \sigma^2\right) \end{aligned}$$

  where $\mathcal{X} := \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ and $\mathcal{Y} := \{y_1, y_2, \ldots, y_n\}$.
- The likelihood and the factors $p(y_n|\boldsymbol{x}_n, \boldsymbol{\theta})$ are Gaussian due to the noise distribution.

Parameter Estimation

- Note that once we have the optimal parameters $\boldsymbol{\theta}^* \in \mathbb{R}^D$, we can predict function values using this parameter estimate. For an arbitrary test input $\boldsymbol{x}_*$ the corresponding distribution of y_* then becomes the following:

$$p(y_*|\boldsymbol{x}_*, \boldsymbol{\theta}) = \mathcal{N}(y_*|\boldsymbol{x}_*^\top \boldsymbol{\theta}^*, \sigma^2)$$

Maximum Likelihood Estimate

- A typically widely used method to find the desired parameters $\boldsymbol{\theta}_{ML}$ is *maximum likelihood estimation*, where we find the parameters that maximize the likelihood.

$$\boldsymbol{\theta}_{ML} = \arg\max_{\boldsymbol{\theta}} p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta})$$

Maximum Likelihood Estimate

- A typically widely used method to find the desired parameters $\boldsymbol{\theta}_{ML}$ is *maximum likelihood estimation*, where we find the parameters that maximize the likelihood.

$$\boldsymbol{\theta}_{ML} = \arg\max_{\boldsymbol{\theta}} p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta})$$

- Important Remark: The likelihood $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$ is not a probability distribution in $\boldsymbol{\theta}$. It is a function of $\boldsymbol{\theta}$ and need not integrate to 1. Note that we compute likelihood for a given $\mathcal{Y}$ and $\mathcal{X}$. When we write $p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta})$, we are talking about the conditional distribution of $\mathcal{Y}$, given a fixed $\mathcal{X}$ and $\boldsymbol{\theta}$. In the case of likelihood, $\boldsymbol{\theta}$ is the variable.

Motivation for the Log Transformation

- Typically differentiating products of functions is much more complex than differentiating the sums of functions.

Motivation for the Log Transformation

- Typically differentiating products of functions is much more complex than differentiating the sums of functions.
- When we want to maximize likelihood, we are trying to maximize the product of several probabilities. This can lead to numerical underflow.

Motivation for the Log Transformation

- Typically differentiating products of functions is much more complex than differentiating the sums of functions.
- When we want to maximize likelihood, we are trying to maximize the product of several probabilities. This can lead to numerical underflow.
- Since logarithm function is monotonic, maximizing the logarithm of a function is equivalent to maximizing the function.

Negative Log Likelihood

- To find the optimal parameters, we minimize the negative log-likelihood as follows

$$-\log p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta}) = -\log \prod_{n=1}^{N} p\left(y_n|\boldsymbol{x}_n, \boldsymbol{\theta}\right) = -\sum_{n=1}^{N} \log p\left(y_n|\boldsymbol{x}_n, \boldsymbol{\theta}\right)$$

Negative Log Likelihood

- To find the optimal parameters, we minimize the negative log-likelihood as follows

$$-\log p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta}) = -\log \prod_{n=1}^{N} p\left(y_n|\boldsymbol{x}_n, \boldsymbol{\theta}\right) = -\sum_{n=1}^{N} \log p\left(y_n|\boldsymbol{x}_n, \boldsymbol{\theta}\right)$$

- Since the likelihood is Gaussian, we have,

$$\log p\left(y_n|\boldsymbol{x}_n, \boldsymbol{\theta}\right) = -\frac{1}{2\sigma^2}\left(y_n - \boldsymbol{x}_n^\top \boldsymbol{\theta}\right)^2 + \text{ const}$$

where the constant is independent of $\boldsymbol{\theta}$.

Negative Log Likelihood

- We therefore get negative log likelihood to be finally,

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &:= \frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(y_n - \boldsymbol{x}_n^\top\boldsymbol{\theta}\right)^2 \\ &= \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}) = \frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}\|^2\end{aligned}$$

  and $\boldsymbol{X} := [\boldsymbol{x}_1, \dots, \boldsymbol{x}_N] \in \mathbb{R}^{N\times D}$ and $\boldsymbol{y} := [y_1, \dots, y_N]^\top \in \mathbb{R}^N$.
- Note that the n^{th} row of $\boldsymbol{X}$ corresponds to training input $\boldsymbol{x}_n$.
- If we minimize the above quantity, we get,

$$\boldsymbol{\theta}_{ML} = (\boldsymbol{X}^\top\boldsymbol{X}^{-1})\boldsymbol{X}^\top\boldsymbol{y}$$

Estimating the noise variance

- Assumption so far: Noise variance σ^2 was known.

Estimating the noise variance

- Assumption so far: Noise variance σ^2 was known.
- Now :Relax this assumption and obtain a maximum likelihood estimator σ^2_{ML} for the noise variance.

Estimating the noise variance

- Assumption so far: Noise variance σ^2 was known.
- Now :Relax this assumption and obtain a maximum likelihood estimator σ^2_{ML} for the noise variance.
- We use the same procedure as above: write down the log-likelihood, compute its derivative with respect to $\sigma^2 > 0$, set it to 0 and obtain the needed estimate.

Estimating the noise variance

- Assumption so far: Noise variance σ^2 was known.
- Now :Relax this assumption and obtain a maximum likelihood estimator σ^2_{ML} for the noise variance.
- We use the same procedure as above: write down the log-likelihood, compute its derivative with respect to $\sigma^2 > 0$, set it to 0 and obtain the needed estimate.
- Final Result:

$$\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(y_n - \boldsymbol{x}_n^\top \boldsymbol{\theta})^2$$

Maximum A Posteriori Estimation

- MLE is prone to overfitting.

Maximum A Posteriori Estimation

- MLE is prone to overfitting.
- Need to mitigate the effects of huge parameter values. How to do this?

Maximum A Posteriori Estimation

- MLE is prone to overfitting.
- Need to mitigate the effects of huge parameter values. How to do this?
- Answer: We place a prior $p(\boldsymbol{\theta})$ on the parameters.

Maximum A Posteriori Estimation

- MLE is prone to overfitting.
- Need to mitigate the effects of huge parameter values. How to do this?
- Answer: We place a prior $p(\boldsymbol{\theta})$ on the parameters.
- Example: Gaussian prior $p(\theta) = \mathcal{N}(0, 1)$ on a parameter which we expect to lie in the interval [-2, 2].

Maximum A Posteriori Estimation

- MLE is prone to overfitting.
- Need to mitigate the effects of huge parameter values. How to do this?
- Answer: We place a prior $p(\boldsymbol{\theta})$ on the parameters.
- Example: Gaussian prior $p(\theta) = \mathcal{N}(0, 1)$ on a parameter which we expect to lie in the interval [-2, 2].
- Once we have a dataset $\mathcal{X}, \mathcal{Y}$, instead of maximizing the likelihood, we seek parameters to maximize the posterior distribution $p(\boldsymbol{\theta}|\mathcal{X}, \mathcal{Y})$.

Maximum A Posteriori Estimation

- From Bayes Theorem, we have

$$p(\boldsymbol{\theta}|\mathcal{X}, \mathcal{Y}) = \frac{p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{Y}|\mathcal{X})}$$

Maximum A Posteriori Estimation

- From Bayes Theorem, we have

$$p(\boldsymbol{\theta}|\mathcal{X}, \mathcal{Y}) = \frac{p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{Y}|\mathcal{X})}$$

- Use the prior distribution $\mathcal{N}(0, b^2 I_n)$

Maximum A Posteriori Estimation

- From Bayes Theorem, we have

$$p(\boldsymbol{\theta}|\mathcal{X}, \mathcal{Y}) = \frac{p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{Y}|\mathcal{X})}$$

- Use the prior distribution $\mathcal{N}(0, b^2 I_n)$
- Draw covariance matrix

Maximum A Posteriori Estimation

- To find the MAP estimate, we follow the same steps as for MLE, firstly by considerating the log-posterior.

$$\log p(\boldsymbol{\theta}|\mathcal{X}, \mathcal{Y}) = \log p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) + \text{ const}$$

Maximum A Posteriori Estimation

- To find the MAP estimate, we follow the same steps as for MLE, firstly by considerating the log-posterior.

$$\log p(\boldsymbol{\theta}|\mathcal{X}, \mathcal{Y}) = \log p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) + \text{ const}$$

- We now minimze the negative log-posterior with respect to $\boldsymbol{\theta}$ to find $\boldsymbol{\theta}_{MAP}$

Maximum A Posteriori Estimation

- To find the MAP estimate, we follow the same steps as for MLE, firstly by considerating the log-posterior.

$$\log p(\boldsymbol{\theta}|\mathcal{X}, \mathcal{Y}) = \log p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) + \text{ const}$$

- We now minimze the negative log-posterior with respect to $\boldsymbol{\theta}$ to find $\boldsymbol{\theta}_{MAP}$
- We have,

$$\boldsymbol{\theta}_{\mathrm{MAP}} \in \arg\min_{\boldsymbol{\theta}} \{-\log p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta}) - \log p(\boldsymbol{\theta})\}$$

Maximum A Posteriori Estimation

- We have

$$\boldsymbol{\theta}_{\mathrm{MAP}} \in \arg\min_{\boldsymbol{\theta}} \{-\log p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta}) - \log p(\boldsymbol{\theta})\}$$

Maximum A Posteriori Estimation

- We have

$$\boldsymbol{\theta}_{\mathrm{MAP}} \in \arg\min_{\boldsymbol{\theta}} \{-\log p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta}) - \log p(\boldsymbol{\theta})\}$$

- Now computing the gradient with respect to $\boldsymbol{\theta}$, we have

$$-\frac{\mathrm{d} \log p(\boldsymbol{\theta}|\mathcal{X}, \mathcal{Y})}{\mathrm{d}\boldsymbol{\theta}} = -\frac{\mathrm{d} \log p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} - \frac{\mathrm{d} \log p(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}}$$

Maximum A Posteriori Estimation

- We have

$$\boldsymbol{\theta}_{\mathrm{MAP}} \in \arg\min_{\boldsymbol{\theta}} \{-\log p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta}) - \log p(\boldsymbol{\theta})\}$$

- Now computing the gradient with respect to $\boldsymbol{\theta}$, we have

$$-\frac{\mathrm{d}\log p(\boldsymbol{\theta}|\mathcal{X}, \mathcal{Y})}{\mathrm{d}\boldsymbol{\theta}} = -\frac{\mathrm{d}\log p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} - \frac{\mathrm{d}\log p(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}}$$

- Using the conjugate Gaussian Prior $p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, b^2 \boldsymbol{I})$ on the parameters $\boldsymbol{\theta}$, we get the negative log posterior as follows:

$$-\log p(\boldsymbol{\theta}|\mathcal{X}, \mathcal{Y}) = \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) + \frac{1}{2b^2}\boldsymbol{\theta}^\top\boldsymbol{\theta} + \text{const}$$

Maximum A Posteriori Estimation

- We have

$$\boldsymbol{\theta}_{\mathrm{MAP}} \in \arg\min_{\boldsymbol{\theta}} \{-\log p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta}) - \log p(\boldsymbol{\theta})\}$$

- Now computing the gradient with respect to $\boldsymbol{\theta}$, we have

$$-\frac{\mathrm{d} \log p(\boldsymbol{\theta}|\mathcal{X}, \mathcal{Y})}{\mathrm{d}\boldsymbol{\theta}} = -\frac{\mathrm{d} \log p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} - \frac{\mathrm{d} \log p(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}}$$

- Using the conjugate Gaussian Prior $p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, b^2 \boldsymbol{I})$ on the parameters $\boldsymbol{\theta}$, we get the negative log posterior as follows:

$$-\log p(\boldsymbol{\theta}|\mathcal{X}, \mathcal{Y}) = \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) + \frac{1}{2b^2}\boldsymbol{\theta}^\top \boldsymbol{\theta} + \text{ const}$$

- If we minimize the above quanitity, we get,

$$\boldsymbol{\theta}_{MAP} = \left(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\sigma^2}{b^2}\boldsymbol{I}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

Maximum A Posteriori Estimation

- In the below example, we place a Gaussian prior $p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ on the parameters $\boldsymbol{\theta}$ and determine the MAP estimates. For the lower order polynomial the effect of the prior is not as pronounced as it is in the case of the higher order polynomial and keeps the polynomial relatively smooth in the second case.

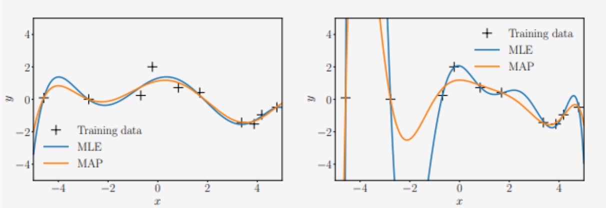


Figure 3: Polynomial Regression and MAP Estimates. Degree 6 and 8 respectively for Figures (a) and (b).

Maximum A Posteriori Estimation

- In the below example, we place a Gaussian prior $p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ on the parameters $\boldsymbol{\theta}$ and determine the MAP estimates. For the lower order polynomial the effect of the prior is not as pronounced as it is in the case of the higher order polynomial and the prior keeps the second polynomial relatively smooth.

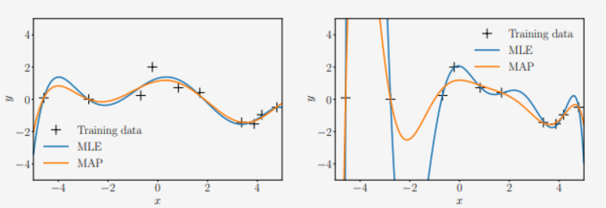


Figure 4: Polynomial Regression and MAP Estimates. Degree 6 and 8 respectively for Figures (a) and (b).

Bayesian Linear Regression

- In Bayesian Linear Regression, we consider the following model:

$$\text{Prior} \;: p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{m_0}, \boldsymbol{S}_0)$$

$$\text{Likelihood} \;: p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{x}^\top\boldsymbol{\theta}, \sigma^2)$$

Bayesian Linear Regression

- In Bayesian Linear Regression, we consider the following model:

$$\text{Prior} \; : p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{m_0}, \boldsymbol{S}_0)$$

$$\text{Likelihood} \; : p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{x}^\top\boldsymbol{\theta}, \sigma^2)$$

- As a PGM, we can represent it as follows:

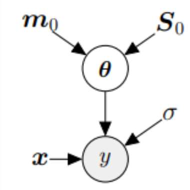


Figure 5: Graphical Model for Bayesian Linear Regression

Bayesian Linear Regression

- The full probabilistic model, i.e., the joint distribution of observed and unobserved random variables, y and $\boldsymbol{\theta}$, respectively, is

$$p(y, \boldsymbol{\theta}|\boldsymbol{x}) = p(y|\boldsymbol{x}, \boldsymbol{\theta})p(\boldsymbol{\theta})$$

Bayesian Linear Regression

- The full probabilistic model, i.e., the joint distribution of observed and unobserved random variables, y and $\boldsymbol{\theta}$, respectively, is

$$p(y, \boldsymbol{\theta} | \boldsymbol{x}) = p(y | \boldsymbol{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta})$$

- The posterior distribution in this case is given by,

$$p(\boldsymbol{\theta} | \mathcal{X}, \mathcal{Y}) = \frac{p(\mathcal{Y} | \mathcal{X}, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathcal{Y} | \mathcal{X})}$$

Bayesian Linear Regression

- The full probabilistic model, i.e., the joint distribution of observed and unobserved random variables, y and $\boldsymbol{\theta}$, respectively, is

$$p(y, \boldsymbol{\theta}|\boldsymbol{x}) = p(y|\boldsymbol{x}, \boldsymbol{\theta})p(\boldsymbol{\theta})$$

- The posterior distribution in this case is given by,

$$p(\boldsymbol{\theta}|\mathcal{X}, \mathcal{Y}) = \frac{p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{Y}|\mathcal{X})}$$

- The denominator above is called as the marginal likelihood or evidence, which ensures that the posterior is normalized and is independent of the parameters. An alternative way of writing the denominator is,

$$p(\mathcal{Y}|\mathcal{X}) = \int p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$$

Parameter Posterior

- The parameter posterior can be computed in closed form as follows:

$$
\begin{aligned}
p(\boldsymbol{\theta}|\mathcal{X}, \mathcal{Y}) &= \mathcal{N}\left(\boldsymbol{\theta}|\boldsymbol{m}_N, \boldsymbol{S}_N\right) \\
\boldsymbol{S}_N &= \left(\boldsymbol{S}_0^{-1} + \sigma^{-2}\mathbf{X}^\top \boldsymbol{X}\right)^{-1} \\
\boldsymbol{m}_N &= \boldsymbol{S}_N\left(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \sigma^{-2}\boldsymbol{X}^\top \boldsymbol{y}\right)
\end{aligned}
$$

Parameter Posterior

- The parameter posterior can be computed in closed form as follows:

$$p(\boldsymbol{\theta}|\mathcal{X},\mathcal{Y}) = \mathcal{N}\left(\boldsymbol{\theta}|\boldsymbol{m}_N,\boldsymbol{S}_N\right)$$
$$\boldsymbol{S}_N = \left(\boldsymbol{S}_0^{-1} + \sigma^{-2}\mathbf{X}^\top\boldsymbol{X}\right)^{-1}$$
$$\boldsymbol{m}_N = \boldsymbol{S}_N\left(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \sigma^{-2}\boldsymbol{X}^\top\boldsymbol{y}\right)$$

- The above posterior follows from:

$$\text{Posterior} \quad p(\boldsymbol{\theta}|\mathcal{X},\mathcal{Y}) = \frac{p(\mathcal{Y}|\mathcal{X},\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{Y}|\mathcal{X})}$$
$$\text{Likelihood} \quad p(\mathcal{Y}|\mathcal{X},\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{\theta},\sigma^2\boldsymbol{I}\right)$$
$$\text{Prior} \quad p(\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\theta}|\boldsymbol{m}_0,\boldsymbol{S}_0\right)$$

Posterior Predictions

- The predictive distribution of y_*, at a test input $\boldsymbol{x}_*$ using the parameter prior $p(\boldsymbol{\theta})$ is computed as follows.

$$\begin{aligned}
p\left(y_* \mid \mathcal{X}, \mathcal{Y}, \boldsymbol{x}_*\right) &= \int p\left(y_* \mid \boldsymbol{x}_*, \boldsymbol{\theta}\right) p(\boldsymbol{\theta} \mid \mathcal{X}, \mathcal{Y}) \mathrm{d}\boldsymbol{\theta} \\
&= \int \mathcal{N}\left(y_* \mid \boldsymbol{x}_*^\top \boldsymbol{\theta}, \sigma^2\right) \mathcal{N}\left(\boldsymbol{\theta} \mid \boldsymbol{m}_N, \boldsymbol{S}_N\right) \mathrm{d}\boldsymbol{\theta} \\
&= \mathcal{N}\left(y_* \mid \boldsymbol{x}_*^\top \boldsymbol{m}_N, \boldsymbol{x}_*^\top \boldsymbol{S}_N \boldsymbol{x}_* + \sigma^2\right)
\end{aligned}$$

Bayesian Linear Regression Analysis

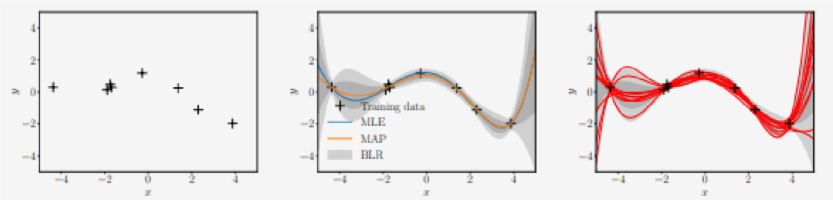
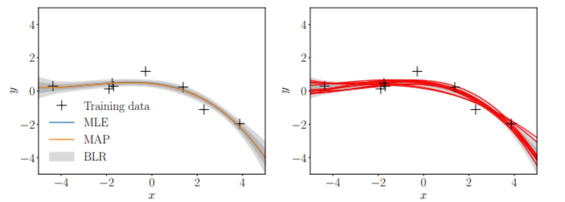
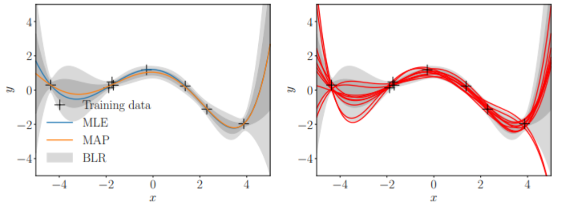


Figure 6: Bayesian linear regression and posterior over functions. (a) training data; (b) posterior distribution over functions; different shades correspond to different confidence intervals (c) Samples from the posterior over functions.

Bayesian Linear Regression Analysis

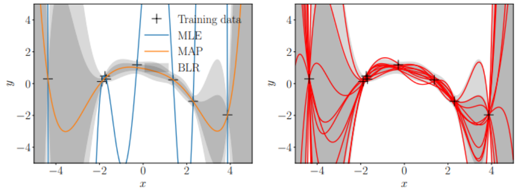


(a) Posterior distribution for polynomials of degree $M = 3$ (left) and samples from the posterior over functions (right).



(b) Posterior distribution for polynomials of degree $M = 5$ (left) and samples from the posterior over functions (right).

Bayesian Linear Regression Analysis



(c) Posterior distribution for polynomials of degree $M = 7$ (left) and samples from the posterior over functions (right).

Figure 7: Left panels: The mean of the Bayesian linear regression model coincides with the MAP estimate. The predictive uncertainty is the sum of the noise term and the posterior parameter uncertainty, which depends on the location of the test input. Right panels: sampled functions from the posterior distribution.