Bayesian Machine Learning

Instructor: Tim G. J. Rudner

Homework 1
Due: Tuesday September 13, 11:59pm via NYU Brightspace

Show all steps, and any code used to answer the questions.

1. Suppose we have data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, and n is the total number of training points. Assume we want to learn the regression model

$$y = ax + \epsilon_x \,, \tag{1}$$

where ϵ_x is independent zero mean Gaussian noise with variance σ^2: $\epsilon_x \sim \mathcal{N}(0, \sigma^2)$.

(a) (2 marks): Let $\mathbf{y} = (y_1, \dots, y_n)^\top$ and $X = \{x_i\}_{i=1}^n$. Derive the log likelihood for the whole training set, $\log p(\mathbf{y}|X, a, \sigma^2)$.

The log likelihood for our data is defined as the probability for seeing all of our observed data points. As our data is i.i.d., we can calculate our log-likelihood (LL) as the product of the probability of seeing each data point.

The probability of seeing any one data point, $\log p(y_i|x_i, a, \sigma^2)$ is defined as follows:

$$\mathcal{N}(y_i, ax_i, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - ax_i)^2}{2\sigma^2}\right)$$

Switching the likelihood to the log form:

$$log\left(\mathcal{N}(y_i, ax_i, \sigma^2)\right) = -\frac{1}{2}log(2\pi) - \frac{1}{2}log(\sigma^2) - \frac{(y_i - ax_i)^2}{2\sigma^2}$$

Doing so for all data points in our data set yields:

$$\log p(\mathbf{y}|X, a, \sigma^2) = -\frac{N}{2}log(2\pi) - \frac{N}{2}log(\sigma^2) - \sum_{i=1}^{n}\left(\frac{(y - ax)^2}{2\sigma^2}\right)$$

Lastly, using matrix vector format for our data, we can re-express our derivation of LL in a compact way:

$$log\left(\mathcal{N}(\mathbf{y}, Xa, \sigma^2)\right) = -\frac{N}{2}log(2\pi) - \frac{N}{2}log(\sigma^2) - \frac{||Xa - \boldsymbol{y}||^2}{2\sigma^2}$$

(b) (2 marks): Given data $\mathcal{D} = \{(4, 21), (9, 59), (7, 25), (15, 127)\}$, find the maximum likelihood solutions for a and σ^2.

We firstly observe that maximizing the log likelihood is the same as minimizing the negative log likelihood (NLL). For sake of convenience, we will elect to go the NLL route. Our expression is:

$$log\left(\mathcal{N}(\mathbf{y}, Xa, \sigma^2)\right) = -\frac{N}{2}log(2\pi) - \frac{N}{2}log(\sigma^2) - \frac{||Xa - \boldsymbol{y}||^2}{2\sigma^2}$$

Differentiating with respect to a, we will disregard the terms that do not contain a. Setting the resulting expression to 0 we have:

$$\begin{aligned} NLL &= \frac{1}{2\sigma^2}(a^T X^T X a - 2a^T X^T \mathbf{y} + \mathbf{y}^T \mathbf{y}) \\ \frac{\partial NLL}{\partial a} &= \frac{1}{2\sigma^2}(2X^T X a - 2X^T \mathbf{y}) = 0 \end{aligned} \tag{2}$$

We can distribute the fraction, and ignore the σ^2 as it does not have an effect on our search of an a that minimizes our NLL. Solving for a:

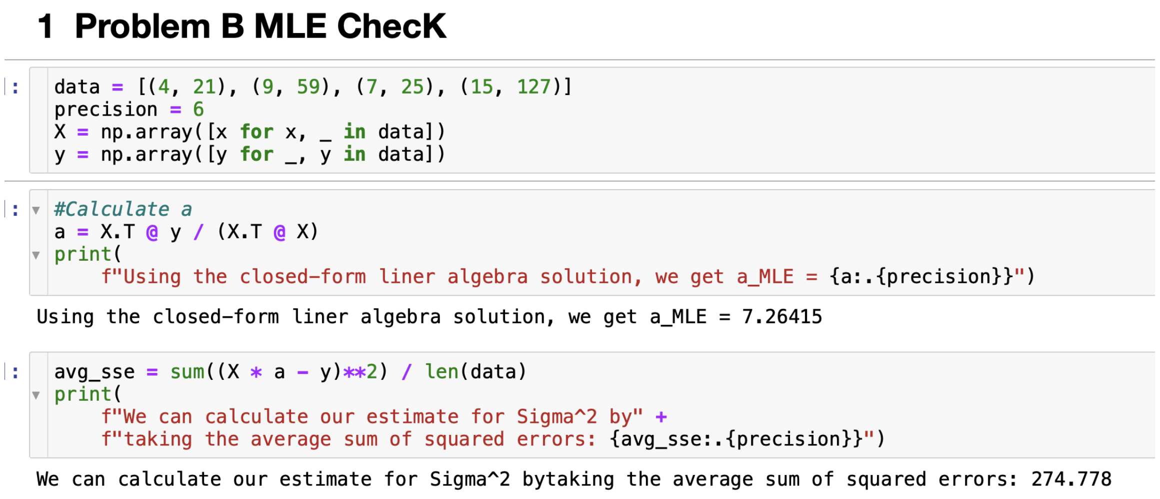
$$\begin{aligned} 0 &= (X^T X a - X^T \mathbf{y}) \\ X^T X a &= X^T \mathbf{y} \\ a_{MLE} &= (X^T X)^{-1} X^T \mathbf{y} \end{aligned} \tag{3}$$

Note that since $X, y \in \mathbb{R}^{n\times 1}$ and $a \in R$, we don't really have a matrix to invert. Our closed form solution more resembles the inner product of X and y over the inner product of X with itself, i.e. $a_{MLE} = \frac{X^T y}{X^T X}$. We can calculate and arrive at $a_{MLE} = 7.26$

Repeating the procedure for σ^2, we take the derivative with respect to σ^2 and set the result to 0.

$$\begin{aligned} NLL &= -\frac{N}{2}log(\sigma^2) - \frac{1}{2\sigma^2}(||Xa - \mathbf{y}||^2) \\ \frac{\partial NLL}{\partial \sigma^2} &= -\frac{N}{2\sigma^2} + \frac{||Xa - \mathbf{y}||^2}{2(\sigma^2)^2} = 0 \\ 2N \times (\sigma^2)^2 &= ||Xa - \mathbf{y}||^2 \times 2\sigma^2 \\ \sigma^2_{MLE} &= \frac{||Xa - \mathbf{y}||^2}{N} \end{aligned} \tag{4}$$

We can calculate our estimator by using a_{MLE} that we calculated earlier, arriving at $SSE = ||Xa_{MLE} - \mathbf{y}||^2 = 1099.133$ and $\sigma^2_{MLE} = \frac{SSE}{4} = 274.788$



1 Problem B MLE CheckK

```
data = [(4, 21), (9, 59), (7, 25), (15, 127)]
precision = 6
X = np.array([x for x, _ in data])
y = np.array([y for _, y in data])
```

```
#Calculate a
a = X.T @ y / (X.T @ X)
print(
    f"Using the closed-form liner algebra solution, we get a_MLE = {a:.{precision}}")
```

```
Using the closed-form liner algebra solution, we get a_MLE = 7.26415
```

```
avg_sse = sum((X * a - y)**2) / len(data)
print(
    f"We can calculate our estimate for Sigma^2 by" +
    f"taking the average sum of squared errors: {avg_sse:.{precision}}")
```

```
We can calculate our estimate for Sigma^2 bytaking the average sum of squared errors: 274.778
```

(c) (2 marks): Suppose we instead consider the regression model

$$x = by + \epsilon\,. \tag{5}$$

Is the maximum likelihood solution for $b = \frac{1}{a}$? Explain why or why not – with derivations if necessary.

Using the same approach we used in problem b) we can take the derivative with respect to b. Doing so, and comparing our derivation with the MLE derivation for a

$$b_{MLE} = \frac{Y^T\mathbf{x}}{Y^TY} \quad a_{MLE} = \frac{X^T\mathbf{y}}{X^TX}$$

As both $X, \mathbf{y} \in \mathbb{R}^{n\times 1}$, we have a dot product in the denominator. We know from the linearity of the dot product, that $X^Ty = \langle X, y\rangle = \langle y, X\rangle = y^TX$, so the numerators are the same, just the denominators are different.

Lets explore a situation where $b = \frac{1}{a}$, we would then have:

$$\begin{aligned} b &= \frac{1}{a} \\ \frac{Y^T\mathbf{x}}{Y^TY} &= \frac{X^TX}{X^T\mathbf{y}} \end{aligned} \tag{6}$$

We can see that for this equivalence to hold, we would need $X^Ty = X^TX = Y^TY$, aka x and y having a 1:1 linear relationship. While this situation is possible, it is not case given our data, as we have data points $x \neq y$.

(d) (2 marks): Suppose we place a prior distribution on a such that $p(a|\gamma^2) = \mathcal{N}(0, \gamma^2)$. Use the sum and product rules of probability to write down the *marginal likelihood* of the data, $p(\mathbf{y}|X, \sigma^2, \gamma^2)$, conditioned only on X, σ^2, γ^2.

Using the sum rule of probability, we can rewrite our marginal likelihood as the integration over a joint distribution on a:

$$p(\mathbf{y}|X, \sigma^2, \gamma^2) = \int p(\mathbf{y}, a|X, \sigma^2, \gamma^2)\mathbf{da}$$

Then using the product rule, we can split apart our joint probability into the product of two probabilities:

$$Marginal\ Likelihood = \int p(\mathbf{y}, a|X, \sigma^2, \gamma^2)\mathbf{da} = \int p(\mathbf{y}|a, X, \sigma^2, \gamma^2)p(a|\gamma^2)\mathbf{da}$$

Note that the resulting equation in the integral is our likelihood function multiplied by the probability density of its corresponding a given our prior on a. We can substitute in our identities and manipulate the expression to show that it results in another Gaussian distribution:

$$\begin{aligned}\int p(\mathbf{y}|a, X, \sigma^2, \gamma^2)p(a|\gamma^2)\mathbf{da} &= \int \mathcal{N}(\mathbf{y}, Xa, \sigma^2)\mathcal{N}(a, 0, \gamma^2) \\ &= \int \frac{N+M}{\sqrt{2\pi}} \times \frac{N}{\sqrt{\sigma^2}} \times \frac{M}{\sqrt{\gamma^2}} \times e^{-\frac{1}{2}\times\left(\frac{||Xa-y||^2}{\sigma^2}+\frac{||a||^2}{\gamma^2}\right)} \\ &= \frac{N+M}{\sqrt{2\pi}} \times \frac{N}{\sqrt{\sigma^2}} \times \frac{M}{\sqrt{\gamma^2}} \int e^{-\frac{1}{2}\times\left(\frac{||Xa-y||^2}{\sigma^2}+\frac{||a||^2}{\gamma^2}\right)}\end{aligned} \tag{7}$$

Evaluating the term inside the integral and exponent, we can expand then term to yield:

$$e^{-\frac{1}{2}\times\left(\frac{||Xa-y||^2}{\sigma^2}+\frac{||a||^2}{\gamma^2}\right)} = \frac{a^TX^TXa - 2a^TX^T\mathbf{y} + \mathbf{y}^T\mathbf{y}}{\sigma^2} + \frac{a^Ta}{\gamma^2}$$

Recongizing we will need to complete the square in the exponent to evaluate the integral, we remove the $\frac{\mathbf{y}^T\mathbf{y}}{\sigma^2}$ scalar from the integral, and move the squared terms together. We set $M = (\frac{X^TX}{\sigma^2} + I\frac{1}{\gamma^2})$ (a positive definite invertible matrix), and $b = \frac{M^{-1}X^Ty}{\sigma^2}$. We then get:

$$\begin{aligned}\frac{a^TX^TXa - 2a^TX^T\mathbf{y}}{\sigma^2} + \frac{a^Ta}{\gamma^2} &= a^T(\frac{X^TX}{\sigma^2} + I\frac{1}{\gamma^2})a + \frac{2a^TX^Ty}{\sigma^2} \\ &= a^TMa - 2a^TMb \\ &= (a-b)^TM(a-b) - b^TMb\end{aligned} \tag{8}$$

In our exponent, we now have:

$$\int e^{-\frac{1}{2}\times(a-b)M(a-b)+b^TMb}$$

However, we can remove the constant, b^TMb, which does not depend on a, outside of our integral. Evaluating our expression and taking the log now yields the following:

$$
\begin{aligned}
Evidence &= \frac{N}{\sqrt{2\pi\sigma^2}} \times \frac{M}{\sqrt{2\pi\gamma^2}} \times \exp\left(-\frac{1}{2}\left(\frac{||\mathbf{y}||^2}{\sigma^2} + b^TMb\right)\right)\int \exp(-\frac{1}{2} \times \left((a-b)^TM(a-b)\right)) \\
LogEvidence &= -\frac{N}{2}log(2\pi) - \frac{M}{2}log(\gamma^2) - \frac{N}{2}log(\sigma^2) - \frac{1}{2}log(|M|) - E(m_n)
\end{aligned}
\tag{9}
$$

Where:

$$
E(m_n) = \frac{1}{2\sigma^2}||y - X^Tb||^2 + \frac{1}{2\sigma^2}b^Tb
$$

(e) (2 marks): Without explicitly using the sum and product rules, derive $p(\mathbf{y}|X, \sigma^2, \gamma^2)$, by considering the properties of Gaussian distributions and finding expectations and covariances. This expression should look different than your answer to the previous question. Comment on the differences in computational complexity. **Bonus (1 mark)**: show that both representations in (d) and (e) are mathematically equivalent.

Starting from equation 1, we have:

$$
y = ax + \epsilon = \mathcal{N}(a|0, \gamma^2)x + \mathcal{N}(0, \sigma^2)
$$

Noting that our equation is the sum of two Gaussian, we know from Bishop that the sum of two Gaussian's is Gaussian's, yielding our marginal likelihood. We can derive the parameters that define $y \sim \mathcal{N}$, (its mean and variance), by taking the expectation of the Gaussian's that define it. Using linearity of expectation, we can remove the nested x from the calculation of the first moment of $a \sim \mathcal{N}$

$$
\mathbf{E}(y) = \mathbf{E}(ax) + \mathbf{E}(\epsilon) = x\mathbf{E}(a) + \mathbf{E}(\epsilon) = 0 + 0 = 0
$$

We can calculate our variance as the summation of two variances that define the Gaussian's that compose y. We can do this as the distributions are independent, so the variances are additive:

$$
\mathbf{E}(y^2) = Var(ax) + Var(\epsilon) = x^2Var(a) + Var(\epsilon) = x^2\sigma^2 + \gamma^2
$$

Generalizing to a covariance matrix in a multidimensional space:

$$
Cov(\mathbf{y}) = \Sigma = X^T\sigma^2X + \gamma^2I
$$

where $\Sigma, X, I, \sigma^2 \in \mathbb{R}^{d\times d}$.

We now have the two quantities that parameterize our distribution: $p(y|X, \sigma^2, \gamma^2) \sim \mathcal{N}(y; 0, X^T\sigma^2X + \gamma^2I)$

Or, alternatively, as $X, y \in \mathbb{R}^{n\times 1}$ we have $p(y|\sigma^2, \gamma^2) \sim \mathcal{N}(y; 0, x^2\sigma^2 + \gamma^2)$

Unfortunately, I'm now seeing that after comparing D and E, having mean 0 for our resulting gaussian doesn't make sense. As I'm writing this right before the assingment is due, I will leave the above included and say that the real shortcut is using the formulas given in Bishop, which would require much less computation than our approach in problem d, as there are no integrals to solve, or "squares to complete". Shortcut as follows:

When integrating over the product of two Gaussians, the result will be a gaussian of the form:

$$\mathcal{N}((y|\sigma^2,\gamma^2), m_n, M)$$

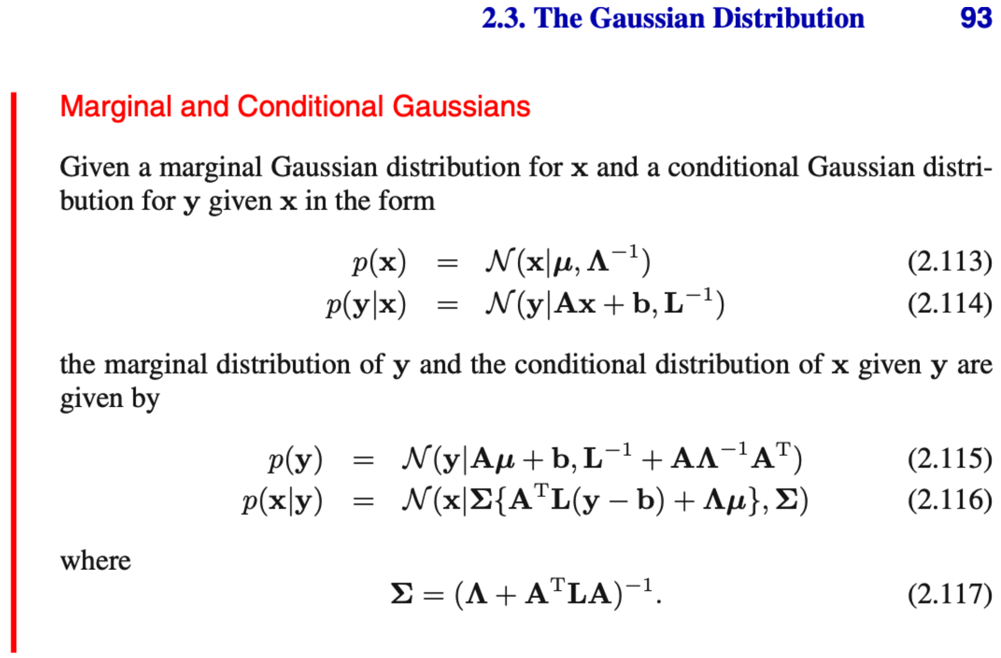
Where

$$M = (\frac{X^T X}{\sigma^2} + I\frac{1}{\gamma^2}) \qquad b = \frac{M^{-1}X^T y}{\sigma^2}$$

And

$$E(m_n) = \frac{1}{2\sigma^2}||y - X^T b||^2 + \frac{1}{2\sigma^2}b^T b$$

Or alternatively, use the formulas from Bishop here:



2.3. The Gaussian Distribution 93

Marginal and Conditional Gaussians

Given a marginal Gaussian distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$ in the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \qquad (2.113)$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x}+\mathbf{b}, \mathbf{L}^{-1}) \qquad (2.114)$$

the marginal distribution of $\mathbf{y}$ and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ are given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu}+\mathbf{b}, \mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}) \qquad (2.115)$$

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y}-\mathbf{b})+\boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) \qquad (2.116)$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}. \qquad (2.117)$$

(f) (2 marks): What are the maximum marginal likelihood solutions $\hat{\sigma}^2 = \mathrm{argmax}_{\sigma^2} p(\mathbf{y}|X, \sigma^2, \gamma^2)$ and $\hat{\gamma}^2 = \mathrm{argmax}_{\gamma^2} p(\mathbf{y}|X, \sigma^2, \gamma^2)$?

We can calculate the log likelihood of our marginal in the following way:

$$
\begin{aligned}
logp(y|\gamma^2,\sigma^2) &= log\int p(\mathbf{y}|a,X,\sigma^2,\gamma^2)p(a|\gamma^2)\mathbf{da} \\
&= -\frac{N}{2}log(2\pi) - \frac{M}{2}log(\gamma^2) - \frac{N}{2}log(\sigma^2) - \frac{1}{2}log|A| - E(m_N) \\
&= -\frac{4}{2}log(2\pi) - \frac{1}{2}log(\gamma^2) - \frac{4}{2}log(\sigma^2) - \frac{1}{2}log(||X||^2\sigma^2+\gamma^2) - \frac{||y||^2}{2\sigma^2} - b^TMb
\end{aligned}
\tag{10}
$$

We can take the derivative with respect to γ^2 and set our expression to 0. (For sake of not breaking my wrists typing latek, I'm combining a few algebra steps into 1)

$$
\begin{aligned}
&\frac{\partial logp(y|\gamma^2,\sigma^2}{\partial\gamma^2} = -(\sigma^2+\gamma^2X^TX)^{-1}X^TX + \frac{||X^Ty||^2}{\sigma^2} + \frac{\partial}{\partial\gamma^2}\frac{\gamma^2}{\sigma^2+\alpha^2X^TX} \\
&(\sigma^2+\alpha^2X^TX) = ||X^Ty||^2
\end{aligned}
\tag{11}
$$

Repeating the process for σ^2:

$$
\begin{aligned}
&\frac{\partial logp(y|\gamma^2,\sigma^2}{\partial\sigma^2} = \frac{(n-1)}{\sigma^2} + \frac{1}{\sigma^2+\gamma^2X^TX} - \frac{||y||^2}{(\sigma^2)^2} - \gamma^2(X^Ty)^2\frac{\partial}{\partial\sigma^2}\frac{(\sigma^2+\gamma^2X^TX)^{-1}}{\sigma^2} \\
&\sigma^2(n-1) = ||y||^2 - \frac{||X^Ty||}{X^TX}
\end{aligned}
\tag{12}
$$

Using the identity we found for sigma, you can plug it in for what we calculated for gamma and now we have:

$$\sigma^2_{MLE} = \frac{1}{N-1}\left(||y||^2 - \frac{||X^Ty||^2}{X^TX}\right) = 366.371$$

$$\gamma^2_{MLE} = \frac{N}{N-1}\left(\frac{X^Ty}{X^TX}\right)^2 - \frac{1}{N-1}\frac{||y||^2}{X^TX} = 51.7804$$

2 Problem F Check

In [27]:

```
xty = X.T@y
xtx = sum(X**2)
yty = sum(y**2)
gamma_mle = ((4/3) * (xty/xtx)**2) - ((1/3) * (yty/xtx))
sigma_mle = (1/3) * (yty - (xty**2/xtx))
```

In [40]:

```
print(f"MLE Estimates as follows: Sigma^2 MLE: {sigma_mle:.{precision}}, Gamma^2 MLE: {gamma_mle:.{precision}}"
```

```
MLE Estimates as follows: Sigma^2 MLE: 366.371, Gamma^2 MLE: 51.7804
```

(g) (2 marks): Derive the predictive distribution for $p(y_*|x_*, \hat{\sigma}^2, \hat{\gamma}^2, \mathcal{D})$ for any arbitrary test point x_*, where $y_* = y(x_*)$.

We can express the predictive distribution for $p(y_*|x_*, \hat{\sigma}^2, \hat{\gamma}^2, \mathcal{D})$ as the following integral using sum and product rules of probability:

$$p(y_*|x_*, \hat{\sigma}^2, \hat{\gamma}^2, \mathcal{D}) = \int p(y_*|x_*, a, \hat{\sigma}^2) p(a|D, \sigma^2, \hat{\gamma}^2)\mathbf{da}$$

Substituting in the Gaussian distribution that define each density:

$$p(y_*|x_*, \hat{\sigma}^2, \hat{\gamma}^2, \mathcal{D}) = \int \mathcal{N}(y(x_*), ax, \hat{\sigma}^2)\mathcal{N}(a, b, M^{-1})\mathbf{da}$$

Using Gaussian identities we covered in part e) we can avoid calculating the integral and instead express the predictive distribution as:

$$p(y_*|x_*, \hat{\sigma}^2, \hat{\gamma}^2, \mathcal{D}) = \mathcal{N}(y(x_*), b^T x, \hat{\sigma}^2 + x_*^T M^{-1} x_*)$$

With

$$M = (\frac{X^T X}{\sigma^2} + I\frac{1}{\gamma^2}) \qquad b = \frac{M^{-1}X^T y}{\sigma^2}$$

.

(h) (2 marks): For the dataset $\mathcal{D}$ in (b), give the predictive mean $\mathbb{E}[y_*|x_*, \hat{\sigma}^2, \hat{\gamma}^2, \mathcal{D}]$ and predictive variance $\text{var}(y_*|x_*, \hat{\sigma}^2, \hat{\gamma}^2, \mathcal{D})$ for $x_* = 14$.

We can use the predictive distribution that we calculated in the previous problem:

$$p(y_*|x_*, \hat{\sigma}^2, \hat{\gamma}^2, \mathcal{D}) = \mathcal{N}(y(x_*), b^T x, \hat{\sigma}^2 + x_*^T M^{-1} x_*)$$

Where $x_* = 14$ $\hat{\sigma^2} = 366.371$ $\hat{\gamma^2} = 51.7804$ from part f), and M and b as we have defined previously. We then have:

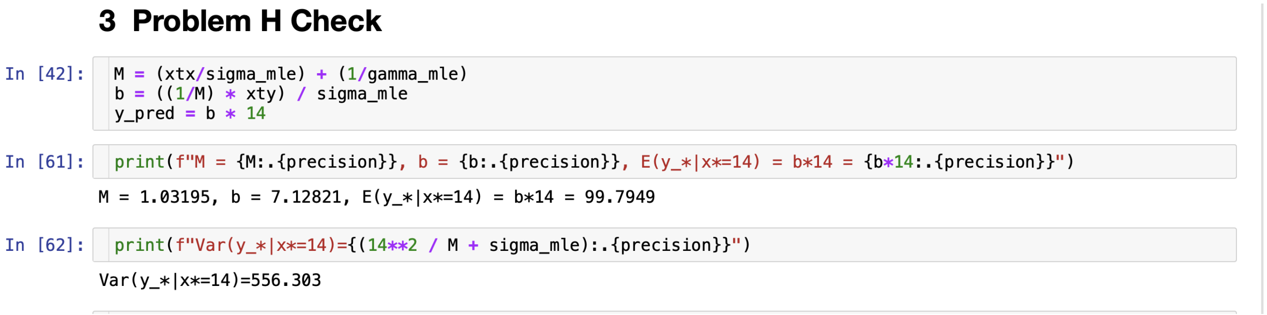
$$M^{-1} = 0.96904 \qquad and \qquad b = 7.12821$$

Evaluating the expectation of the predictive we then have:

$$\mathbb{E}[y_*|x_*, \hat{\sigma}^2, \hat{\gamma}^2, \mathcal{D}] = b^T x = 14 * 7.12821 = 99.7949$$

Evaluating the variance of the predictive we then have:

$$Var[y_*|x_*, \hat{\sigma}^2, \hat{\gamma}^2, \mathcal{D}] = \hat{\sigma}^2 + x_*^T M^{-1} x_* = 366.371 + \left(.96904 \times 14^2\right) = 556.30332$$



3 Problem H Check

```
In [42]:  M = (xtx/sigma_mle) + (1/gamma_mle)
          b = ((1/M) * xty) / sigma_mle
          y_pred = b * 14
```

```
In [61]:  print(f"M = {M:.{precision}}, b = {b:.{precision}}, E(y_*|x*=14) = b*14 = {b*14:.{precision}}")

          M = 1.03195, b = 7.12821, E(y_*|x*=14) = b*14 = 99.7949
```

```
In [62]:  print(f"Var(y_*|x*=14)={(14**2 / M + sigma_mle):.{precision}}")

          Var(y_*|x*=14)=556.303
```

(i) (2 marks): Suppose we replace x in Eq. (1) with $g(x, w)$, where g is a non-linear function parametrized by w, and $w \sim \mathcal{N}(0, \lambda^2)$: e.g., $g(x, w) = \cos(wx)$. Can you write down an analytic expression for $p(\mathbf{y}|w, X, \sigma^2, \gamma^2)$? How about $p(\mathbf{y}|X, \sigma^2, \gamma^2, \lambda^2)$? Justify your answers.

I'm having a hard time understanding this question – If we plan on getting rid of a, and we just predict off of $g(x, w)$, then its nonsense. If we are referring to $g(x, w)$ like we would some non linear basis kernel function, and have the expression:

$$y = g(w, x)a + \epsilon$$

Then its just kernalization, which is fair game as we can define an arbitrary non-linear mapping function $g(w, x)$ (normally referred to as $\psi(x)$) which mapped/transformed our x input to a higher dimensional space, then we could parameterize the higher dimensional output with a weight vector a, in our case yielding a linear combination of the parameters i.e.

$$p(y|x, \sigma^2) \sim \mathcal{N}(y, \psi(x, w)^T a, \sigma^2)$$

We could then derive an expression for $p(\mathbf{y}|w, X, \sigma^2, \gamma^2)$ by simply replacing every occurrence of X with $\psi(X, w)$.

$$p(\mathbf{y}|w, X, \sigma^2, \gamma^2) = -\frac{N}{2}log(2\pi) - \frac{M}{2}log(\gamma^2) - \frac{N}{2}log(\sigma^2) - \frac{1}{2}log(|M|) - E(m_n) \quad (13)$$

Where:

$$E(m_n) = \frac{1}{2\sigma^2}||y - \psi(X, w)^T b||^2 + \frac{1}{2\sigma^2}b^T b$$

$$M = (\frac{\psi(X, w)^T \psi(X, w)}{\sigma^2} + I\frac{1}{\gamma^2}) \qquad b = \frac{M^{-1}\psi(X, w)^T y}{\sigma^2}$$

.

Placing a prior on our variable w, we could derive the marginal distribution of y across its relevant parameters by integrating over w, so long as the prior is also Gaussian and therefore conjugate.

$$p(\mathbf{y}|X, \sigma^2, \gamma^2, \lambda^2) = \int\int \mathcal{N}(y, a^T\psi(x, w), \sigma^2)\mathcal{N}(a, 0, \gamma^2)\mathcal{N}(w, 0, \lambda^2)\ \mathbf{da\ dw}$$

However, if missing the a, or not picking a Gaussian prior, there's no guarantee this will work:

$$p(y|x, \sigma^2) \sim \mathcal{N}(y, cos(x, w), \sigma^2)$$

There would be no guarantee of being able to write an analytic expression as we wouldn't be using a conjugate prior, and the resulting integral could potentially be unsolvable or fail to yield a valid probability density (i.e. negative densities).