Assignment 1

Date 01 Feb 2026

Question 1

(a) In the formulation of a linear regression model, ϵ_n is the noise term, which has no straight relationship with the data yet. Given that $\epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, and this is a random variable which adds to $w^T \cdot x$ term in order to get a particular y value from the model, we easily formulate $y_n \overset{iid}{\sim} \mathcal{N}(w^T x, \sigma^2)$, with mean $= 0 + w^T x$ and unchanged variance σ^2 in an additive transformation. Now, we can bridge the relationship between model's noises, weights, and features, with a Gaussian (Normal) distribution: $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(y - w^T x)^2}{2\sigma^2}\right)$, w.r.t. x

y is derived w.r.t. x, w, therefore the joint likelihood is $\prod_{i=1}^{N} p(y_i \mid x_i, w)$

To maximize the likelihood, one could use an integral:

$$\int p(y_n \mid x_n, w)\, dx\, dw$$

But this is intractable (computationally inefficient), so, alternatively, we could take log: $\sum_{i=1}^{N} \ln p(y_i \mid x_i, w)$

Subs. into $p(x)$, we have:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \ln\left(\exp\left(-\frac{\sum_{i=1}^{N}(y_i - x_i \cdot w^T)^2}{2\sigma^2}\right)\right)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \left(-\frac{\sum_{i=1}^{N}(y_i - w^T \cdot x_i)^2}{2\sigma^2}\right)$$

We all know that, σ^2 is partially known. The goal is to maximize the evidence $p(x)$ in order to maximize the likelihood, jointly from $\prod_{i=1}^{N} p(y_i \mid x_i, w)$.

Constants have no impact on the maximization, so clearly we are finding the maximum of $-\sum_{i=1}^{N}(y_i - w^T x_i)^2$

Date 01 Feb 2026

Question 1 (contd.)

(a) which is to find $p(x)$ with the minimum $(y_i - w^T x_i)$ terms. With all given data samples y_i and x_i, where $i = 1, \ldots, N$ and N is the dataset size, we are, hence, finding the optimal

$$\hat{w} = \arg\min_{w} \sum_{n=1}^{N} (y_n - w^T \cdot x_n)^2 \quad \square$$

(b) For Bayesian Linear Regression, we generally assume random variables follow an independent identical distribution (i. i. d.), which means, in our case $\varepsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, we have the same variance σ^2 for all observed / measured data samples (x_i, y_i).

From the notation $\mathcal{N}(0, \sigma^2)$, a random variable must follow a Gaussian distribution, in our assumptions, where Normal posterior $\propto$ Normal prior $\times$ Normal likelihood.

And as the name suggests, the function must be linear, which means the degree of polynomial = 1, in this form: $y = mx + c$.

Assignment 1

Date 01 Feb 2026

Question 2

(a) Given $\Sigma_N = (\beta X^T X + \lambda I_D)^{-1}$

$$= \frac{1}{\beta X^T X + \lambda I_D}$$

In a formulation from a covariance of a posterior, β and λ are hyperparameters, which are constants in our problem. I_D is an identity matrix which is also a constant.

X is a matrix with data samples. Given that $X^T X$ is increasing, the value of the denominator (which is a sum of 2 products) must also grow. But as it is in the denominator Σ_N value shrinks as $X^T X$ grows, and N grows, where the number of data samples increases. Σ_N is non-proportional to N and $X^T X$.

It is valid in pratical machine learning too. Covariance is the variance of a multi-variate distribution. Variance depicts the variability of a model's prediction with the true value, which is aleatorically impacted by the noise. Epistemically adding samples (terms) to the model's formula makes it more confident about its understanding on the data and thus, less variability.

Date 01 Feb 2026

Question 3

(a) Let w be the weight in the function to derive y^* from x^*. $\beta = \frac{1}{\sigma^2}$ which is the precision.

$$p(y^* | x^*, X, y, \beta, \lambda)$$

$$= \int p(y^* | x^*, w, \beta)\, p(w | X, y, \beta, \lambda)\, dw$$

By marginalizing w from the posterior, we know

$$p(y^* | x^*) = \int p(y^* | x^*, w)\, p(w | X, y)\, dw$$

We can also generalize the distribution:

$$p(w | X, y, \beta, \lambda) = \mathcal{N}(w; \mu_N, \Sigma_N)$$

where μ_N is the mean and Σ_N is the covariance after learning N samples from the prior, for the marginal of w. In this case, we can also derive

$$p(w) = \mathcal{N}(\mu_N, \Sigma_N)$$

Considering how we derive the new label y^* from an existing model with weights as w, we need to know $p(y^* | w)$

Let $A = (x^*)^T$, we can derive:

$$\text{mean } \mu_y = A\mu_N + b$$

$$\text{co-variance } \Sigma_y = \Sigma_{y|x} + A\Sigma_x A^T$$

Subs. this into $p(y^* | x^*)$, we have:

$$p(y^* | x^*) = \mathcal{N}(\mu_y, \Sigma_y)$$

$$= \mathcal{N}(A\mu_N + b,\ \Sigma_{y|x} + A\Sigma_x A^T)$$

A is the Affine Matrix to ensure $y = Ax + b$. for any sample (x, y).

Assignment 1 Date 01 Feb 2026

Question 3 (contd.)

(a) Subs. $A = (x^*)^T$ definition back into $p(y^*|x^*)$

we have: $p(y^*|x^*) = \mathcal{N}(x^{*T}\mu + b, \Sigma_{y|x} + x^{*T}\Sigma_N x^*)$

We know the precision β depicting how confident the true label is within the posterior, whereas the co-variance (in multi-variate situations) depicting how uncertain the true label is within that range. So, we can define the proportionality: $\beta = \frac{1}{\sigma^2}$, in other words,

$$p(y^*|x^*) = \mathcal{N}(x^*\mu + b, \beta^{-1} + x^{*T}\Sigma_N x^*)$$
$$= p(y^*|x^*, X, y, \beta, \lambda).$$

(b) Aleatoric uncertainty is those that are un-controllable. Apparently, it's the noise of a model, which is shown by variance σ^2. As we know from (a), precision $\beta = \frac{1}{\sigma^2}$, we can deduce that β^{-1} is aleatoric

Epistemic uncertainty is those that are controllable. We can say those variables that are data-dependent. In our case, it is $x^{*T}\Sigma_N x^*$.

Meanwhile, b is a constant. But if we are not sure what the value of b is the best, then we lack knowledge in defining b. We could define another prior distribution for b. Since its uncertainty is still manageable, it is also epistemic.

Date 01 Feb 2026

Question 3 (contd.)

(c) $\beta = \frac{1}{\sigma^2}$, where σ^2 is the variance of a distribution. With this formula, it is clearly that β is inversely proportional to σ^2. In practical machine learning, while the variance (or covariance Σ_N in a multi-variate distribution) shows the spread, or to be precise, the uncertainty, of a distribution, β measures the data precision. So, β^{-1} is the variance.

(d) Because it is a multi-variate normal (Gaussian) distribution, we need to construct a matrix in a quadratic form which satisfies the definition of variance : $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}$, where x_i is a data sample, $\bar{x}$ is the sample mean, and N is the total number of samples. Co-variance is defined based on multiple feature variables :

$$cov(x_i, x_j) = E[(x_i - E(x_i))(x_j - E(x_j))]$$

From $\mathcal{N}(\mu_N^T x^*, \beta^{-1} + x^{*T} \Sigma_N x^*)$, we need the multiplication of x^* with itself in transpose in order to define the covariance's matrix.

(e) Maximum Likelihood Estimation (MLE) estimates an optimal parameter : $\hat{\theta} = \arg\max_\theta \log p(X \mid \theta)$

Maximum A Posteri (MAP) estimates it with a regularizer : $\hat{\theta} = \arg\min_\theta [-\log p(X|\theta) - \log p(\theta)]$

$\Rightarrow \hat{\theta} = \arg\max_\theta (\log p(X|\theta) + \log p(\theta))$

Assignment 1

Date 01 Feb 2026

Question 3 (contd.)

(e) MLE simply maximizes the log-likelihood with differentiation, as the likelihood formula is convex where the global optimum is easily derived from its first derivative.
MAP, on the other hand, balances data with prior. By shrinking the prior while maximizing the posterior $p(D|\theta)p(\theta) = p(\theta|D)$, it pulls the estimation towards the mean.

In our Bayesian Linear Regression, we make a balance between aleatoric and epistemic uncertainties, from the variance: $\beta^{-1} + x^{*T}\Sigma_N x^*$.
In order to determine the best parameters, i.e., w in (a), we need to reduce Σ_N, as β^{-1} is not in our scope of control (aleatoric), and in the multiplication term, we have to increase $x^{*T} \cdot x^*$ and shrink Σ_N in order to reduce overall variance. It is evident that adding more data samples (i.e., increasing N) could help reducing epistemic uncertainty, and thus, the variance. Hence, this is the best.

In addition, when we explore the $\arg\max_\theta$ components in MLE and MAP formulations respectively, it is only approximating from a point estimate, while Bayesian Linear regression caters for the entire distribution of samples. Hence, this is more robust.

Date 01 Feb 2026

Question 4

(a) Let $y = f(x)$ be the function that we are visualizing, which follows $\mathcal{N}(\mu, 2^2)$. We can plot the uncertainty in a data space or with the posterior.

In a data space, we plot the input x versus the predictive output y. When there are almost no data points, there are many lines, trying to robustly cover all possible values, especially outliers. And it is the prior $p(w)$, where w is the weights and this refers to the model's original understanding. As we have more data points, the function tries to fit as many points as possible. The variance shrinks as the function shifts its focus, from robustness to data precision. But when there are overly many data samples, although the variance tends to 0, it might overfit existing samples and be unable to fit new ones. Here, variance shows the change of an uncertainty band.

If we plot the posterior $p(w|X)$, in the relationship of different weight terms, we can clearly see circular or oval shapes, which is the uncertainty band, visualized from the variance. With almost no data samples, it is close to the prior, so the band will look like a circle with radius = variance = 2^2 = 4 as given. It aims to robustly cover all outliers. As the sample size increases, the size (i.e., the radius) shrinks, as it focuses on precision, fitting data points.

Assignment 1 Date 02 Feb 2026

Question 4 (contd.)

(b) Given the posterior predictive distribution for y^* at a new input x^*: $p(y^* | x_*, X, y, \beta, \lambda) = \mathcal{N}(\mu_N^T x^*, \beta^{-1} + x^{*T} \Sigma_N x^*)$, the variance of the entire Gaussian distribution depends on $\beta^{-1} + x^{*T} \Sigma_N x^*$. In the addition, β^{-1} is independent of the data x^*. We can treat it an aleatoric uncertainty, as a constant. Variance $\propto x^{*T} \Sigma_N x^*$. As described in (a), we know that when sample size increases, variance decreases, so they are inversely proportional. Therefore, in this product, we can deduce variance $\propto \Sigma_N$, it is the covariance term, showing the epistemic uncertainty.

Question 5

(a) Entropy-based uncertainty sampling.

(b) Since the posterior predictive variance is epistemically controlled by the sample size and Σ_N the co-variance, we need a method that effectively reduces the spread of data. As an active learning strategy, uncertainty sampling itself does not simply receive samples passively, but it chooses a x^* which is closest to the learned decision boundary. It helps getting data points closer to one another, and thus, reduces the spread, with a simple subtraction: $x^* = \arg\min_x P(\hat{y}_1 | x, \theta) - P(\hat{y}_2 | x, \theta)$. Introducing an entropy makes it more tractable in a practical machine learning use case: $x^* = \arg\max_x - \sum_i P(y_i | x, \theta) \log P(y_i | x, \theta)$, with logarithm rather than integral.