Assignment 1 - ELG 5218 Uncertainty Evaluation in Engineering Measurements and Machine Learning

Miodrag Bolic

2026-01-25

Each item (a, b, ...) in the questions carries a 1-point weight unless otherwise specified. There are 29 points in total. You need 25 points to get 100%, the remaining 4 are the bonus.

1 Linear regression - short questions. (Total 17 points)

Q1. Ordinary least squares as maximum likelihood.

Consider the linear regression model

$$y_n = w^\top x_n + \epsilon_n, \quad n = 1, \ldots, N,$$

with $\epsilon_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ for some fixed $\sigma^2 > 0$.

(a) Show that ordinary least squares (OLS) estimation of w, defined as

$$\hat{w}_{\text{OLS}} = \arg\min_w \sum_{n=1}^{N} (y_n - w^\top x_n)^2,$$

is equivalent to the maximum likelihood estimator (MLE) for w under the Gaussian noise assumption above.

(b) State clearly which probabilistic assumptions are required for this equivalence to hold. What are the assumptions for the Bayesian linear regression in general?

Q2. Effect of sample size on posterior covariance.

The posterior covariance for w can be written as

$$\Sigma_N = (\beta X^\top X + \lambda I_D)^{-1}.$$

(a) Explain qualitatively how Σ_N changes as N increases, assuming that $X^\top X$ grows and remains well-conditioned.

Q3. Predictive variance and input dependence.

The posterior predictive distribution for y^* at a new input x^* is

$$p(y^* \mid x^*, X, y, \beta, \lambda) = \mathcal{N}\left(\mu_N^\top x^*,\ \beta^{-1} + x^{*\top} \Sigma_N x^*\right).$$

(a) Derive posterior distribution using theorems from `https://github.com/Health-Devices/Course-Uncertaint blob/master/Lec2/GaussianFormulas.pdf`

(b) What part of the predictive covariance is related to aleatoric and what to epistemic uncertainty?

(c) Explain the interpretation of the term β^{-1}.

(d) Explain the interpretation of the term $x^{*\top}\Sigma_N x^*$ and why it depends on x^*.

(e) Contrast this with plug-in MLE/MAP prediction, where the predictive variance does not depend on x^* in the same way.

Q4. Predictive bands in data-rich vs data-poor regions.

(a) Suppose you plot the predictive mean and mean ± 2 standard deviations as a function of a single scalar input x. Describe the qualitative shape of the uncertainty band in regions where many data points are available versus regions where almost no data have been observed.

(b) Explain how this behavior connects to the geometry of $x^\top \Sigma_N x$.

Q5. Using predictive uncertainty for active learning (advanced - not covered in the class.)

In active learning, instead of labeling a large random dataset upfront, the model adaptively chooses which input points to label next in order to gain as much information as possible per label. A very common family of strategies is uncertainty sampling: always query the point for which the current model is most uncertain about its prediction. Imagine you can query new inputs x^* to label out of a pool of unlabeled samples. Uncertainty Sampling is the simplest and most commonly used query framework (Lewis and Gale, 1994). In this framework, an active learner queries the instances about which it is least certain how to label. Read intro to active learning 1.1, 1.2, 2.3 and 3.1 from Active Learning Literature Survey, by Burr Settles

(a) Propose a simple active learning strategy that uses the posterior predictive variance $\beta^{-1} + x^{*\top}\Sigma_N x^*$ to select the next point x^*.

(b) Explain why this strategy prioritizes informative queries (Informative queries are unlabeled input points whose labels would be maximally helpful for improving the model) in the Bayesian linear regression setting.

Q6. Interpreting Bayesian linear regression + coding. The figure illustrates the process of sequential Bayesian inference for estimating the parameters of a linear regression model of the form

$$p(y \mid x) = \mathcal{N}(y \mid w_0 + w_1 x, \sigma^2).$$

Each row corresponds to a different stage of the inference process after observing $N = 0$, 1, 2, and 100 data points. The left column shows the likelihood function of the most recently observed data point. The middle column displays the posterior distribution over the model parameters (w_0, w_1) after incorporating the first N observations, i.e.,

$$p(w_0, w_1 \mid x_{1:N}, y_{1:N}, \sigma^2).$$

The right column shows samples from the posterior predictive distribution, illustrating how uncertainty in the parameters propagates to predictions.

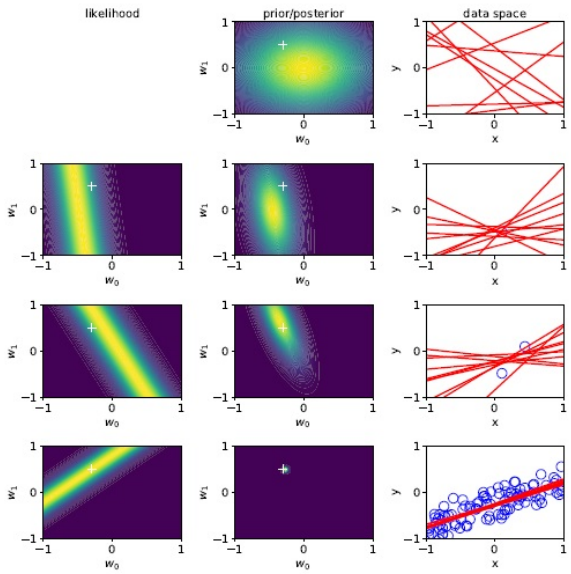

Figure 1: Bayesian linear regression.

In both the likelihood and posterior plots, a white cross marks the true parameter values used to generate the data.

To generate the figure, synthetic training data were sampled from a linear model with known parameters. Visualizations are shown at four stages of the Bayesian update process, corresponding to data indices $N = 0, 1, 2$, and 100.

The true regression parameters used to generate the data were:

$$a_0 = -0.3, \qquad a_1 = 0.5,$$

chosen within the interval $[-1, 1]$ to ensure well-behaved contour plots.

A total of 100 training samples (x, y) were generated. Gaussian noise with standard deviation 0.2 was added to each output value to simulate measurement uncertainty. The Bayesian prior over (w_0, w_1) was a zero-mean isotropic Gaussian with prior precision $\alpha = 2.0$, representing moderate initial uncertainty. The likelihood assumed the same noise standard deviation as the data-generating process.

The parameter settings used were:

Number of training points:	100,
Noise standard deviation:	0.2,
Prior precision (α) :	2.0,
Likelihood standard deviation:	0.2,
Data indices visualized:	0, 1, 2, 100.

(a) (3 points) Implement Python code that reproduces the three-column, four-row figure for Bayesian linear regression with Gaussian prior and noise. Use the parameters specified above.

(b) (2 points) Interpret the following:

i. **Likelihood shape:** classify as parallel/diagonal/isotropic and explain how (x_N, y_N) induces that ridge.

ii. **Posterior center (numeric):** report $(\hat{w}_0, \hat{w}_1) = \mathbf{m}_N$ rounded to 2 decimals.

iii. **Posterior spread & correlation:** uncertainty = high/medium/low; corr(w_0, w_1) = positive/negative/nea zero.

iv. **Convergence:** relative to the white cross, does the posterior mode/mean move closer/unchanged/farther?

2 Notebook-Based Questions (Total 4 points)

This question is related to Part 3 of the notebook Gaussian_Models_Final.ipynb.

Q7. Modifying the sensor noise.

Extend the Part 3 code by changing the temperature sensor noise `sensor_noise` from 1.5 to 0.5 and then to 3.0.

(a) For each value of `sensor_noise`, generate the four plots (posterior mean, precision, learning rate, variance) and describe how the learning dynamics change.

(b) Specifically comment on how the speed of convergence of μ_t and the magnitude of w_t depend on the assumed measurement noise.

Q8. Extending Part 3: Non-stationary environment.

Modify the Part 3 temperature example so that the true temperature slowly drifts over time (e.g., `true_temp_t = 22 + 0.1 * t`).

(a) Implement this change and rerun the online update algorithm without modifying the prior or update formulas. Plot the posterior mean trajectory against the drifting true temperature.

(b) Discuss whether the Bayesian learner can track the drift and what limitations arise from using a "static-parameter" model in a non-stationary environment.

3 Bayesian Analysis of the Exponential Distribution (Total 8 points)

A machine lifetime X is modeled as an exponential random variable with unknown rate parameter θ. The probability density is

$$p(x \mid \theta) = \theta e^{-\theta x}, \qquad x \geq 0,\ \theta > 0.$$

Q9. (Theory – MLE)

(a) Show that the maximum likelihood estimator (MLE) for θ based on independent observations $x_1, \ldots, x_N$ is

$$\hat{\theta} = \frac{1}{\bar{x}}, \qquad \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i.$$

Q10. (Numerical MLE computation)

Suppose we observe the lifetimes (in years) of three independent machines:

$$X_1 = 5, \quad X_2 = 6, \quad X_3 = 4.$$

(a) Compute the MLE of θ using these data.

Q11. (Bayesian inference) An expert specifies an exponential-form prior for θ:

$$p(\theta) = \theta^3 e^{-3\theta}, \qquad \theta > 0.$$

(a) Derive the posterior distribution $p(\theta \mid x_1, \ldots, x_N)$.

(b) Determine whether this prior is conjugate to the exponential likelihood, and justify your answer.

Q12. (Coding Problem – Simulation and Bayesian Updating) Write Python code to do the following:

(a) Simulate $N = \{10, 100\}$ lifetimes from Exponential(θ_0) with true rate $\theta_0 = 0.2$.

(b) Compute and print the MLE $\hat{\theta}$ from the simulated data.

(c) Using the prior

$$p(\theta) = \theta^3 e^{-3\theta},$$

compute the posterior distribution parameters and plot the posterior density. On the same plot, show the likelihood function (up to a constant). and the prior density for comparison.

(d) Discuss in 3–4 sentences whether the posterior is consistent with the theoretical conjugacy behavior you derived in part Q11(a).

Provide your code in a clearly formatted Python block, and produce a figure showing the prior, likelihood, and posterior on the same set of axes.