Signal Processing

Lecture 14: Bayesian Inference

Konstantinos Chatzilygeroudis - costashatz@upatras.gr

Department of Electrical and Computer Engineering
University of Patras

Template made by Panagiotis Papagiannopoulos

Bayesian Inference: Core Idea

Unknown parameters as random variables

$\boldsymbol{\theta}$ is uncertain $\Rightarrow$ treat $\boldsymbol{\theta}$ as a random variable and encode belief with a prior $p(\boldsymbol{\theta})$.

Observations are random too

$$\boldsymbol{Y} \text{ is generated from a model conditioned on } (\boldsymbol{X}, \boldsymbol{\theta}),$$

so we need a probabilistic data model.

Goal: infer $\boldsymbol{\theta}$ from data

$$\text{Given } (\boldsymbol{X}, \boldsymbol{Y}), \quad \text{compute a distribution over } \boldsymbol{\theta}.$$

Why Bayesian?

- Gives *uncertainty* about parameters.
- Naturally incorporates *prior knowledge / regularization*.
- Enables *model comparison* via evidence.

Bayesian Inference: Likelihood

Data model (likelihood)

$$p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}) \quad \text{describes how data is generated given } \boldsymbol{\theta}.$$

Interpretation

- A function of $\boldsymbol{Y}$ for fixed $\boldsymbol{\theta}$: a probability model.
- A function of $\boldsymbol{\theta}$ for fixed data: a measure of *how compatible* $\boldsymbol{\theta}$ is with the observations.

Example (Gaussian noise)

$$y_i = f(x_i; \boldsymbol{\theta}) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$$\Rightarrow \quad p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}) = \prod_{i=1}^{N} \mathcal{N}\left(y_i; f(x_i; \boldsymbol{\theta}), \sigma^2\right).$$

Bayes' Rule and Posterior

Bayes' rule: posterior distribution

$$p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X}) = \frac{p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\boldsymbol{Y} \mid \boldsymbol{X})}.$$

Evidence / Marginal likelihood

$$p(\boldsymbol{Y} \mid \boldsymbol{X}) = \int p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Key message

$$\boxed{\text{Posterior} \propto \text{Likelihood} \times \text{Prior}.}$$

Prior, Likelihood, Posterior, Evidence

Interpretation:

- **Prior** $p(\boldsymbol{\theta})$: knowledge / belief *before* seeing data.
- **Likelihood** $p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta})$: how likely data is under $\boldsymbol{\theta}$.
- **Posterior** $p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X})$: updated belief after seeing data.
- **Evidence** $p(\boldsymbol{Y} \mid \boldsymbol{X})$: normalizes and measures model plausibility.

Intuition

- The likelihood pulls $\boldsymbol{\theta}$ toward values that explain data.
- The prior pulls $\boldsymbol{\theta}$ toward values we believed beforehand.
- The posterior balances the two.

Evidence and Model Comparison

Why do we care about evidence?

$$p(\boldsymbol{Y} \mid \boldsymbol{X}) = \int p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

Interpretation

- It is the probability of observing the data under the model *before* knowing parameters.
- It averages likelihood over the prior: good models explain data *for many plausible parameters*.

Model comparison (Bayes factor) For two models $\mathcal{M}_1, \mathcal{M}_2$:

$$\frac{p(\boldsymbol{Y} \mid \boldsymbol{X}, \mathcal{M}_1)}{p(\boldsymbol{Y} \mid \boldsymbol{X}, \mathcal{M}_2)}$$ compares how well each model explains the data.

Occam's razor effect: complex models are penalized unless strongly supported by data.

Point Estimation as Risk Minimization

Bayesian inference gives a *distribution* over parameters. Often we want a single estimate $\hat{\boldsymbol{\theta}}$.

Decision-theoretic view Choose $\hat{\boldsymbol{\theta}}$ to minimize posterior expected loss:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\tilde{\boldsymbol{\theta}}} \mathbb{E}\Big[\mathcal{L}(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) \;\Big|\; \boldsymbol{Y}, \boldsymbol{X}\Big].$$

Different losses $\mathcal{L}$ lead to different estimators:

- 0–1 loss $\Rightarrow$ MAP.
- Squared error loss $\Rightarrow$ MMSE.
- Absolute error loss $\Rightarrow$ posterior median.

Common Bayesian Estimators – MAP

Given the posterior

$$p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X}) \propto p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta})\, p(\boldsymbol{\theta}),$$

1) MAP (Maximum A Posteriori): posterior mode

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X}) = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta})\, p(\boldsymbol{\theta}).$$

Equivalent minimization form

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\min_{\boldsymbol{\theta}} \Big(-\log p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}) - \log p(\boldsymbol{\theta}) \Big).$$

Interpretation: maximize posterior $\Leftrightarrow$ minimize *data misfit* + *regularization*.

The optimization over the MAP criterion results to the parameters, $\hat{\boldsymbol{\theta}}_{\text{MAP}}$, that are the *posterior distribution's mode value*. In other words, we find *the peak of the posterior probability density function* (in general, we may have multiple peaks!).

MAP, Priors, and Regularization

Example: Gaussian prior

$$p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, \alpha^{-1} I) \quad \Rightarrow \quad -\log p(\boldsymbol{\theta}) \propto \frac{\alpha}{2} \|\boldsymbol{\theta}\|_2^2.$$

With Gaussian noise likelihood:

$$-\log p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}) \propto \frac{1}{2\sigma^2} \|\boldsymbol{Y} - f(\boldsymbol{X}; \boldsymbol{\theta})\|_2^2$$

MAP becomes regularized least squares

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma^2} \|\boldsymbol{Y} - f(\boldsymbol{X}; \boldsymbol{\theta})\|^2 + \frac{\alpha}{2} \|\boldsymbol{\theta}\|^2.$$

So: **priors correspond to regularizers.**

MLE as a Special Case of MAP

If $p(\boldsymbol{\theta}) \propto$ const (flat / uninformative prior), then

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}) = \hat{\boldsymbol{\theta}}_{\text{MLE}}.$$

Interpretation

- MAP uses *prior* + *data*.
- MLE uses *data only*.
- With enough data, MAP and MLE often get close.

Common Bayesian Estimators – MMSE

2) MMSE (Minimum Mean Square Error): posterior mean
Definition (squared-error loss):

$$\hat{\boldsymbol{\theta}}_{\text{MMSE}} = \arg\min_{\tilde{\boldsymbol{\theta}}} \mathbb{E}\Big[\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_2^2 \;\Big|\; \boldsymbol{Y}, \boldsymbol{X} \Big].$$

where $\boldsymbol{\theta}$ is the *true (latent) parameter* treated as a random vector, and $\tilde{\boldsymbol{\theta}}$ is a *deterministic estimate* we choose to minimize posterior risk.

Result: the minimizer is the posterior mean

$$\boxed{\hat{\boldsymbol{\theta}}_{\text{MMSE}} = \mathbb{E}[\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X}] = \int \boldsymbol{\theta}\, p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X})\, d\boldsymbol{\theta}.}$$

Interpretation:

- MMSE averages all plausible parameters, weighted by posterior probability.
- Optimal when we care about *expected squared deviation* from the truth.
- More robust than MAP when the posterior is skewed or multimodal.

Why MMSE Equals the Posterior Mean? (Sketch)

We want to minimize

$$\mathbb{E}\Big[\|\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}\|^2 \mid \boldsymbol{Y},\boldsymbol{X}\Big].$$

Expand around the posterior mean $\boldsymbol{\mu}=\mathbb{E}[\boldsymbol{\theta}\mid \boldsymbol{Y},\boldsymbol{X}]$:

$$\|\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}\|^2=\|(\boldsymbol{\theta}-\boldsymbol{\mu})+(\boldsymbol{\mu}-\tilde{\boldsymbol{\theta}})\|^2=\|\boldsymbol{\theta}-\boldsymbol{\mu}\|^2+\|\boldsymbol{\mu}-\tilde{\boldsymbol{\theta}}\|^2+2(\boldsymbol{\theta}-\boldsymbol{\mu})^\top(\boldsymbol{\mu}-\tilde{\boldsymbol{\theta}}).$$

Take posterior expectation:

$$\mathbb{E}[\boldsymbol{\theta}-\boldsymbol{\mu}\mid \boldsymbol{Y},\boldsymbol{X}]=\mathbf{0} \quad \Rightarrow \quad \text{cross term vanishes.}$$

Thus

$$\mathbb{E}\Big[\|\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}\|^2 \mid \cdot\Big]=\underbrace{\mathbb{E}\big[\|\boldsymbol{\theta}-\boldsymbol{\mu}\|^2 \mid \cdot\big]}_{\text{independent of }\tilde{\boldsymbol{\theta}}}+\|\boldsymbol{\mu}-\tilde{\boldsymbol{\theta}}\|^2,$$

minimized when $\tilde{\boldsymbol{\theta}}=\boldsymbol{\mu}$.

Common Bayesian Estimators – MMAE

3) MMAE (Minimum Mean Absolute Error): posterior median
Definition (absolute-error loss):

$$\hat{\boldsymbol{\theta}}_{\mathrm{MMAE}} = \arg\min_{\tilde{\boldsymbol{\theta}}} \mathbb{E}\Big[\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_1 \;\Big|\; \boldsymbol{Y}, \boldsymbol{X} \Big].$$

where $\boldsymbol{\theta}$ is the *true (latent) parameter* treated as a random vector, and $\tilde{\boldsymbol{\theta}}$ is a *deterministic estimate* we choose to minimize posterior risk.
Result: the minimizer is the posterior median. We assume $\|\cdot\|_1 = \sum_i |\cdot|$,

$$\boxed{\hat{\boldsymbol{\theta}}_{\mathrm{MMAE}} = \big[\mathrm{median}(p(\theta_1 \mid \boldsymbol{Y}, \boldsymbol{X})), \ldots, \mathrm{median}(p(\theta_d \mid \boldsymbol{Y}, \boldsymbol{X}))\big]^\top.}$$

i.e., the estimator is the *componentwise posterior median*.
Interpretation:

- MMAE picks the most "central" posterior value in terms of *absolute deviation*.
- Optimal when we care about *expected absolute error* rather than squared error.
- More robust to outliers/heavy tails than MMSE (mean can be pulled by extremes).
- For skewed posteriors, median and mean can differ substantially.

MAP vs MMSE vs MMAE

MAP:

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X}) \quad \text{(posterior mode)}$$

MMSE:

$$\hat{\boldsymbol{\theta}}_{\text{MMSE}} = \mathbb{E}[\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X}] \quad \text{(posterior mean)}$$

MMAE:

$$\hat{\boldsymbol{\theta}}_{\text{MMAE}} = \arg\min_{\tilde{\boldsymbol{\theta}}} \mathbb{E}\Big[\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_1 \;\Big|\; \boldsymbol{Y}, \boldsymbol{X} \Big] \quad \text{(posterior median)}$$

When are they equal? If $p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X})$ is Gaussian (or any symmetric unimodal pdf):

$$\text{mean} = \text{median} = \text{mode} \quad \Rightarrow \quad \hat{\boldsymbol{\theta}}_{\text{MMSE}} = \hat{\boldsymbol{\theta}}_{\text{MMAE}} = \hat{\boldsymbol{\theta}}_{\text{MAP}}.$$

Practical intuition

- MAP picks the *single most probable* parameter.
- MMSE averages all plausible parameters with posterior weights.
- MMAE selects the *central* posterior value, robust to outliers.
- For skewed / heavy-tailed posteriors:

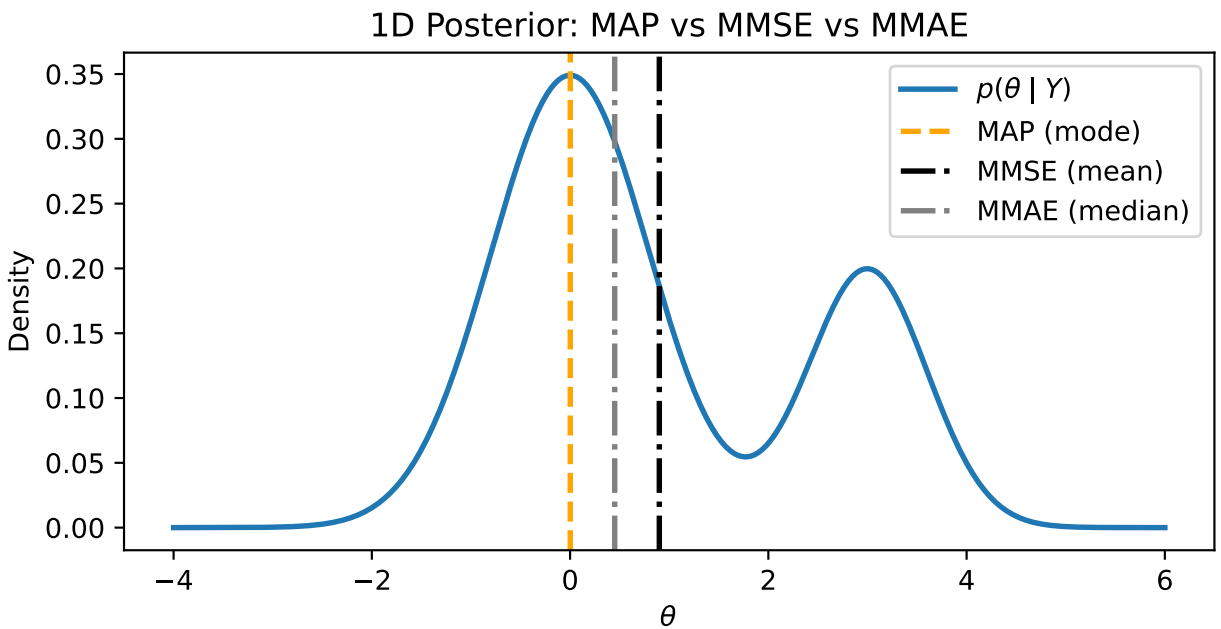
$$\text{mode} \neq \text{median} \neq \text{mean},$$

so the three estimators can differ substantially.

1D Illustration

Consider a posterior $p(\theta \mid Y)$ in 1D.

- **MAP** $\hat{\theta}_{\text{MAP}}$ is at the *peak*.
- **MMSE** $\hat{\theta}_{\text{MMSE}}$ is the *center of mass*.
- **MMAE** $\hat{\theta}_{\text{MMAE}}$ is the *median*.



Key idea: different estimators reflect different optimality criteria.

Maximum Likelihood Estimation (MLE)

Setup: we observe data $\boldsymbol{y}$ generated by a model parameterized by $\boldsymbol{\theta}$ with likelihood

$$p(\boldsymbol{y} \mid \boldsymbol{\theta}).$$

Goal (MLE): choose parameters that make the observed data most probable:

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} \; p(\boldsymbol{y} \mid \boldsymbol{\theta}).$$

Log-likelihood (equivalent objective):

$$\mathcal{L}(\boldsymbol{\theta}) = \log p(\boldsymbol{y} \mid \boldsymbol{\theta}), \qquad \hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} \; \mathcal{L}(\boldsymbol{\theta}).$$

(Log is used because it turns products into sums and is numerically stable.)
If samples are i.i.d.:

$$p(\boldsymbol{y} \mid \boldsymbol{\theta}) = \prod_{k=1}^{N} p\big(y_k \mid \boldsymbol{\theta}\big) \quad \Rightarrow \quad \mathcal{L}(\boldsymbol{\theta}) = \sum_{k=1}^{N} \log p\big(y_k \mid \boldsymbol{\theta}\big).$$

How do we solve it?

- If $\mathcal{L}(\boldsymbol{\theta})$ is concave and simple $\Rightarrow$ closed form.
- Otherwise $\Rightarrow$ iterative optimization (Newton / Gauss–Newton methods / gradient descent).

Key idea: MLE depends on the assumed noise/data distribution in $p(\boldsymbol{y} \mid \boldsymbol{\theta})$.

Maximum Likelihood Estimation: i.i.d. Data

Data: N i.i.d. input–output samples

$$\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N}, \qquad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^{\mathsf{T}} \\ \boldsymbol{x}_2^{\mathsf{T}} \\ \vdots \\ \boldsymbol{x}_N^{\mathsf{T}} \end{bmatrix}, \quad \boldsymbol{Y} = \begin{bmatrix} \boldsymbol{y}_1^{\mathsf{T}} \\ \boldsymbol{y}_2^{\mathsf{T}} \\ \vdots \\ \boldsymbol{y}_N^{\mathsf{T}} \end{bmatrix}.$$

Model: conditional distribution of outputs given inputs and parameters with **i.i.d. assumption:**

$$p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}) = \prod_{i=1}^{N} p\big(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}\big).$$

Log-likelihood: (turn product into sum)

$$\mathcal{L}(\boldsymbol{\theta}) = \log p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}) = \sum_{i=1}^{N} \log p\big(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}\big).$$

MLE objective:

$$\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \arg\max_{\boldsymbol{\theta}} \; \mathcal{L}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \log p\big(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}\big).$$

Note: the form of $p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta})$ depends on the assumed noise / data model.

MLE for Linear Models with Gaussian Noise

Data: N i.i.d. pairs $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ with $\boldsymbol{x}_i \in \mathbb{R}^p, \qquad y_i \in \mathbb{R}$.
Linear model (scalar output):

$$y_i = \boldsymbol{\theta}^{\mathsf{T}} \boldsymbol{x}_i + w_i, \qquad \boldsymbol{\theta} \in \mathbb{R}^p.$$

Gaussian noise assumption:

$$w_i \sim \mathcal{N}(0, \sigma^2) \quad \Rightarrow \quad p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}^{\mathsf{T}} \boldsymbol{x}_i, \sigma^2).$$

Log-likelihood:

$$\mathcal{L}(\boldsymbol{\theta}) = \log p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}) = \sum_{i=1}^N \log p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}).$$

For Gaussian noise:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^N \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \left(y_i - \boldsymbol{\theta}^{\mathsf{T}} \boldsymbol{x}_i\right)^2 \right].$$

MLE (maximize log-likelihood):

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) \iff \underbrace{\arg\min_{\boldsymbol{\theta}} \sum_{i=1}^N \left(y_i - \boldsymbol{\theta}^{\mathsf{T}} \boldsymbol{x}_i\right)^2}_{\text{Linear Least Squares!}}.$$

Bayesian Linear Regression (BLR): Problem Setting

We observe training pairs $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ and assume

$$y_i = \boldsymbol{\theta}^\mathsf{T} \boldsymbol{\phi}(\boldsymbol{x}_i) + w_i, \qquad w_i \sim \mathcal{N}(0, \sigma_w^2).$$

Feature map:

$$\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^P \quad \text{(could be polynomial, Fourier, RBF, etc.)}$$

Matrix form:

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{\theta} + \boldsymbol{w}, \quad \boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \sigma_w^2 \boldsymbol{I}), \quad \boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\phi}(\boldsymbol{x}_1)^\mathsf{T} \\ \vdots \\ \boldsymbol{\phi}(\boldsymbol{x}_N)^\mathsf{T} \end{bmatrix}.$$

Goal: infer $\boldsymbol{\theta}$ and predict y_* with uncertainty.

BLR: Bayesian Viewpoint

Parameters are random variables

$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta}) \quad \text{(encode uncertainty / prior knowledge)}.$$

Gaussian prior (conjugate choice):

$$p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0).$$

Interpretation

- $\boldsymbol{\mu}_0$: prior guess of weights.
- $\boldsymbol{\Sigma}_0$: prior uncertainty / correlation between weights.
- If $\boldsymbol{\Sigma}_0 = \alpha^{-1}\boldsymbol{I}$: isotropic prior $\Rightarrow$ shrink weights toward zero.

BLR: Likelihood

Given $\boldsymbol{\theta}$, the data likelihood is Gaussian:

$$p(\boldsymbol{y} \mid \boldsymbol{\Phi}, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\Phi}\boldsymbol{\theta}, \sigma_w^2 \boldsymbol{I}).$$

Log-likelihood

$$\log p(\boldsymbol{y} \mid \boldsymbol{\Phi}, \boldsymbol{\theta}) = -\frac{N}{2} \log(2\pi\sigma_w^2) - \frac{1}{2\sigma_w^2} \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}\|^2.$$

Connection

$$\text{Gaussian noise} \;\Rightarrow\; \text{Least Squares} \;\Leftrightarrow\; \text{MLE}.$$

BLR: Prior Predictive Distribution (Before Data)

Prior prediction at new input x_*:

$$p(y_* \mid \boldsymbol{x}_*) = \int p(y_* \mid \boldsymbol{x}_*, \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Since (likelihood)

$$p(y_* \mid \boldsymbol{x}_*, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\phi}_*^{\mathsf{T}}\boldsymbol{\theta}, \sigma_w^2), \quad \boldsymbol{\phi}_* \equiv \boldsymbol{\phi}(\boldsymbol{x}_*),$$

introduce the **noiseless latent prediction**

$$z_* \equiv \boldsymbol{\phi}_*^{\mathsf{T}}\boldsymbol{\theta}.$$

Because $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and z_* is a linear function of $\boldsymbol{\theta}$,

$$z_* \sim \mathcal{N}\Big(\boldsymbol{\phi}_*^{\mathsf{T}}\boldsymbol{\mu}_0,\ \boldsymbol{\phi}_*^{\mathsf{T}}\boldsymbol{\Sigma}_0\boldsymbol{\phi}_*\Big).$$

Now $y_* = z_* + w_*$ with $w_* \sim \mathcal{N}(0, \sigma_w^2)$ independent, so the sum of Gaussians is Gaussian:

$$\boxed{p(y_* \mid \boldsymbol{x}_*) = \mathcal{N}\Big(\boldsymbol{\phi}_*^{\mathsf{T}}\boldsymbol{\mu}_0,\ \boldsymbol{\phi}_*^{\mathsf{T}}\boldsymbol{\Sigma}_0\boldsymbol{\phi}_* + \sigma_w^2\Big).}$$

Intuition:

- We predict by averaging over all plausible $\boldsymbol{\theta}$ under the prior.
- Predictive variance = parameter uncertainty + measurement noise.
- No data yet: predictions are driven only by the prior.

BLR: Generating Synthetic Data from the Prior

Prior sampling idea: before observing any data, we can *sample plausible models* from the prior and generate synthetic observations.

Step 1: sample parameters

$$\boldsymbol{\theta}^{(s)} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \qquad s = 1, \ldots, S.$$

Step 2: form noiseless functions

$$z_*^{(s)} = \boldsymbol{\phi}(\boldsymbol{x}_*)^\mathsf{T} \boldsymbol{\theta}^{(s)}.$$

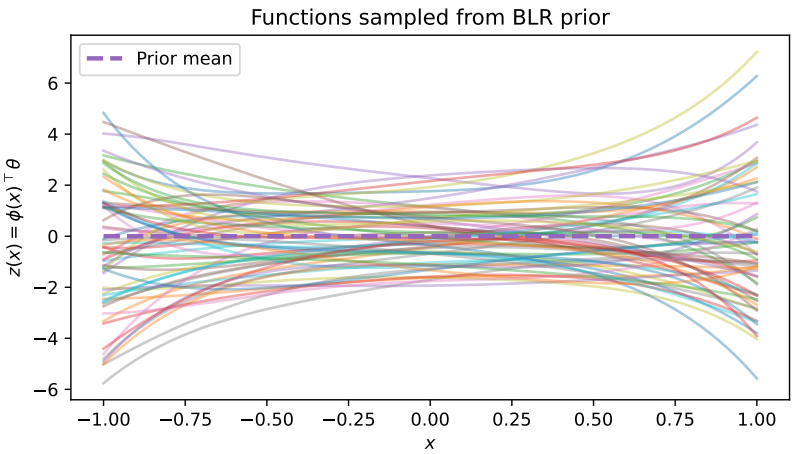
Step 3: add observation noise

$$y_*^{(s)} = z_*^{(s)} + w_*^{(s)}, \qquad w_*^{(s)} \sim \mathcal{N}(0, \sigma_w^2).$$

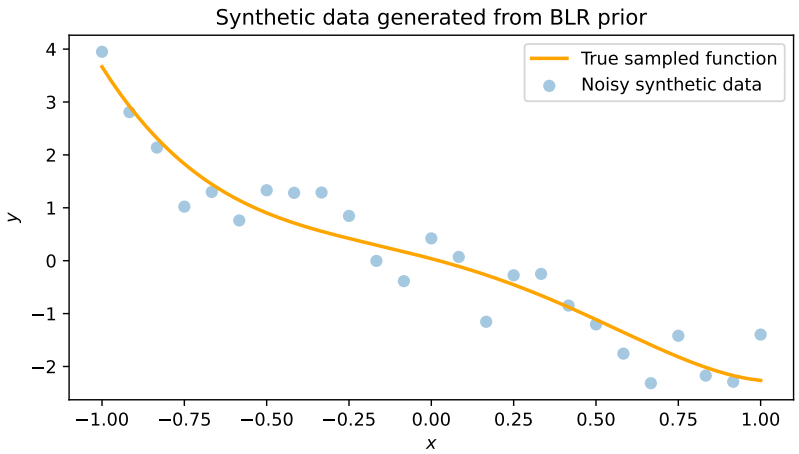
Why useful?

- Visualize what functions the prior considers plausible.
- Create toy datasets for demos / assignments.
- Stress-test inference: does BLR recover $\boldsymbol{\theta}$ under its own assumptions?
- Debug priors/features before collecting real data.

BLR: Prior Sampling and Synthetic Data



Prior function samples



One synthetic dataset

- Left: plausible functions before seeing data.
- Right: sample one function, add noise to generate training points.

BLR: Posterior is Gaussian (Conjugacy)

Posterior

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{\Phi}) \propto p(\boldsymbol{y} \mid \boldsymbol{\Phi}, \boldsymbol{\theta})\, p(\boldsymbol{\theta}).$$

Because both prior and likelihood are Gaussian in $\boldsymbol{\theta}$,

$$\boxed{p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{\Phi}) = \mathcal{N}(\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N).}$$

Meaning: data updates our belief, but keeps a Gaussian form.

BLR: Derivation (Complete the Square)

Start from log posterior up to constants:

$$\log p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{\Phi}) = -\frac{1}{2\sigma_w^2}\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}\|^2 - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_0)^\mathsf{T}\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_0).$$

Expand the quadratic terms in $\boldsymbol{\theta}$:

$$= -\frac{1}{2}\Big[\boldsymbol{\theta}^\mathsf{T}\Big(\boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma_w^2}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\Big)\boldsymbol{\theta} - 2\boldsymbol{\theta}^\mathsf{T}\Big(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \frac{1}{\sigma_w^2}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{y}\Big)\Big].$$

Identify Gaussian form $\Rightarrow$ read off posterior mean/covariance.

$$\text{Let } A = \boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma_w^2}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}, \quad b = \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \frac{1}{\sigma_w^2}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{y}.$$

$$-\frac{1}{2}\Big[\boldsymbol{\theta}^\mathsf{T}A\boldsymbol{\theta} - 2\boldsymbol{\theta}^\mathsf{T}b\Big] = -\frac{1}{2}(\boldsymbol{\theta} - A^{-1}b)^\mathsf{T}A(\boldsymbol{\theta} - A^{-1}b) + \text{const.}$$

$$\Rightarrow\ p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{\Phi}) = \mathcal{N}(\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N), \quad \boldsymbol{\Sigma}_N = A^{-1}, \quad \boldsymbol{\mu}_N = A^{-1}b.$$

BLR: Posterior Mean and Covariance

$$\boxed{p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{\Phi}) = \mathcal{N}(\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N).}$$

$$\boxed{\boldsymbol{\Sigma}_N = \left(\boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma_w^2}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)^{-1}} \qquad \boxed{\boldsymbol{\mu}_N = \boldsymbol{\Sigma}_N\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \frac{1}{\sigma_w^2}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{y}\right)}.$$

Precision form (often cleaner):

$$\boldsymbol{\Sigma}_N^{-1} = \boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma_w^2}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}.$$

Interpretation

- Prior precision + data precision.
- More / cleaner data $\Rightarrow$ larger precision $\Rightarrow$ smaller variance.

BLR vs LS: MAP Connection

MAP estimator (posterior mode):

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{\Phi}).$$

Using the log posterior:

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_w^2} \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}\|^2 + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_0)^{\mathsf{T}} \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_0).$$

Special case $\boldsymbol{\mu}_0 = \mathbf{0},\ \boldsymbol{\Sigma}_0 = \alpha^{-1}\boldsymbol{I}$:

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}\|^2 + \lambda \|\boldsymbol{\theta}\|^2, \quad \lambda = \sigma_w^2 \alpha,$$

i.e. **ridge regression / Tikhonov regularization.**

BLR: Posterior Predictive (After Data)

Predict at x_* after observing data:

$$p(y_* \mid \boldsymbol{x}_*, \boldsymbol{y}, \boldsymbol{\Phi}) = \int p(y_* \mid \boldsymbol{x}_*, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{\Phi})\, d\boldsymbol{\theta}.$$

With

$$p(y_* \mid \boldsymbol{x}_*, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\phi}_*^{\mathsf{T}}\boldsymbol{\theta}, \sigma_w^2), \quad p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{\Phi}) = \mathcal{N}(\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N),$$

we get:

$$\boxed{p(y_* \mid \boldsymbol{x}_*, \boldsymbol{y}, \boldsymbol{\Phi}) = \mathcal{N}\Big(\boldsymbol{\phi}_*^{\mathsf{T}}\boldsymbol{\mu}_N,\ \boldsymbol{\phi}_*^{\mathsf{T}}\boldsymbol{\Sigma}_N\boldsymbol{\phi}_* + \sigma_w^2\Big).}$$

Posterior Predictive: Mean vs Variance

Predictive mean:

$$\mathbb{E}[y_* \mid \cdot] = \boldsymbol{\phi}_*^{\mathsf{T}} \boldsymbol{\mu}_N \quad \text{(use posterior mean weights)}.$$

Predictive variance:

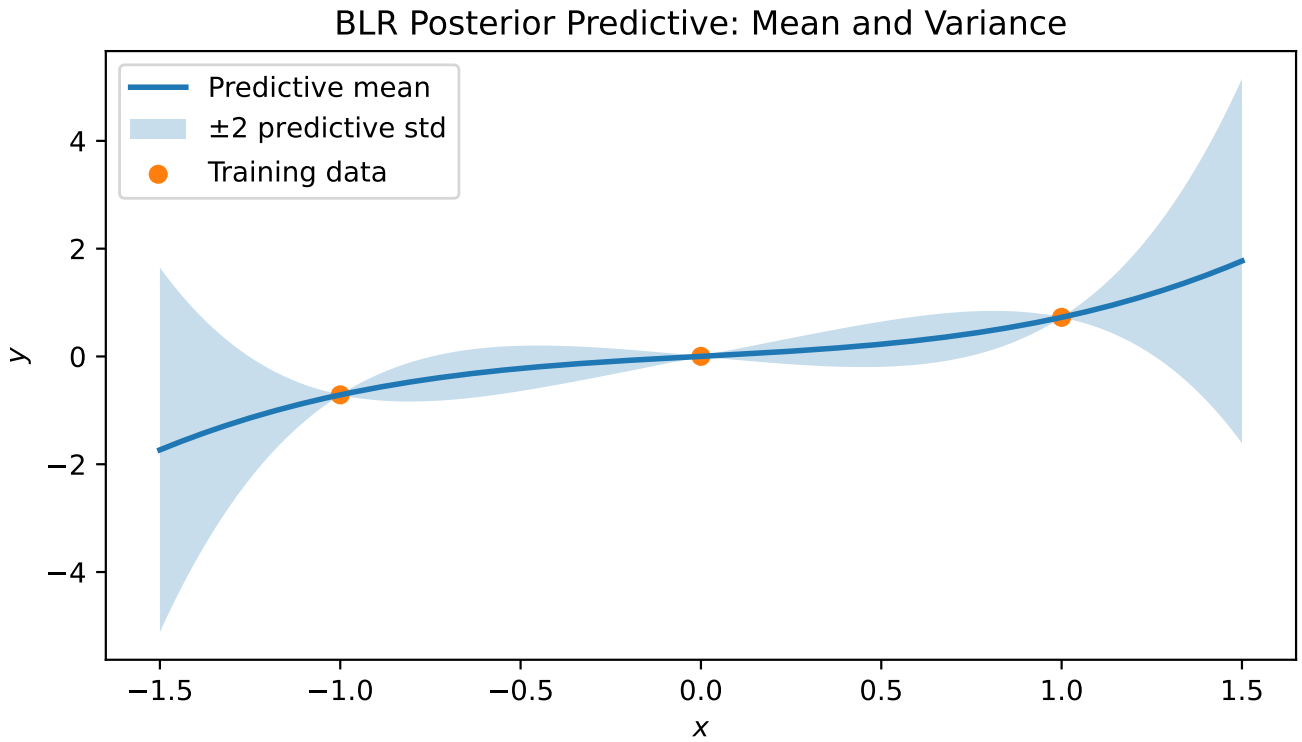
$$\mathrm{Var}(y_* \mid \cdot) = \underbrace{\boldsymbol{\phi}_*^{\mathsf{T}} \boldsymbol{\Sigma}_N \boldsymbol{\phi}_*}_{\text{parameter uncertainty}} + \underbrace{\sigma_w^2}_{\text{measurement noise}}.$$

Key messages

- Even with zero noise ($\sigma_w^2 \to 0$), uncertainty remains if $\boldsymbol{\Sigma}_N \neq 0$.
- As $N \to \infty$, $\boldsymbol{\Sigma}_N \to \mathbf{0}$: uncertainty bands tighten.

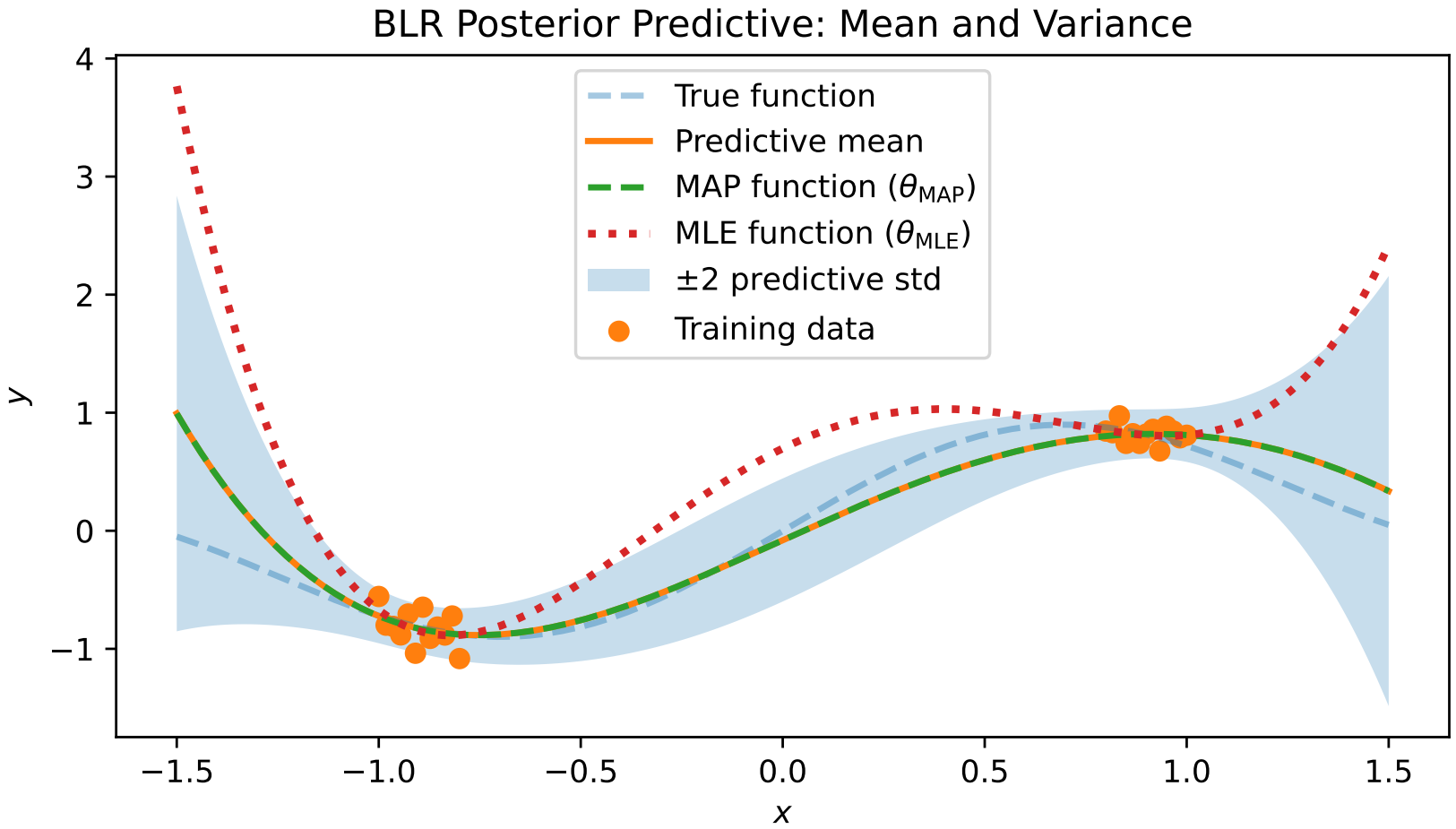
Posterior Predictive: Mean vs Variance (Example)

Example: BLR with polynomial features on noisy 1D data.



- Solid line: predictive mean $\mathbb{E}[y_* \mid \cdot] = \phi_*^\top \mu_N$.
- Shaded band: ± 2 predictive std, $\sqrt{\phi_*^\top \Sigma_N \phi_* + \sigma_w^2}$.
- Uncertainty is small near data, larger far away (parameter uncertainty).

Posterior Predictive: BLR vs MLE vs MAP



BLR: Sequential / Recursive Update

BLR can be updated one data point at a time. Assume current posterior after k samples:

$$p(\boldsymbol{\theta} \mid \mathcal{D}_k) = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

Given new sample $(\boldsymbol{x}_{k+1}, y_{k+1})$ with feature $\boldsymbol{\phi}_{k+1}$:

$$\boldsymbol{\Sigma}_{k+1}^{-1} = \boldsymbol{\Sigma}_k^{-1} + \frac{1}{\sigma_w^2}\boldsymbol{\phi}_{k+1}\boldsymbol{\phi}_{k+1}^{\mathsf{T}},$$

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\Sigma}_{k+1}\left(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k + \frac{1}{\sigma_w^2}\boldsymbol{\phi}_{k+1}y_{k+1}\right).$$

BLR: What We Gain vs LS/MLE

- A *distribution* over parameters (uncertainty on $\boldsymbol{\theta}$).
- Built-in regularization via the prior.
- Predictive uncertainty:

$$p(y_* \mid \boldsymbol{x}_*, \boldsymbol{y}) = \int p(y_* \mid \boldsymbol{x}_*, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{y})\, d\boldsymbol{\theta}.$$

- With enough data, posterior concentrates and approaches the MLE.

Takeaway: BLR $=$ $\underbrace{\text{LS/MLE} + \text{prior knowledge}}_{\text{MAP}}$ $+$ uncertainty quantification.

Thank you

- **Any Questions?**
- **Office Hours**:
  - **Tue & Thu** *(09:00-11:00)*
  - 24/7 by email (`costashatz@upatras.gr`, subject: *ECE_SP_AM*)
- **Material and Announcements**

Laboratory of Automation & Robotics