Imperial College London

Lecture 4: Bayesian linear regression

Introduction to machine learning

Kevin Webster

Department of Mathematics
Imperial College London

Outline

Outline

Linear regression: review

We briefly review the setting of linear regression.

We are given a dataset consisting of N input points

$$\mathbf{x} = (x_1, x_2, \ldots, x_N), \qquad x_i \in \mathbb{R}^d \quad \forall i$$

and N corresponding output values

$$\mathbf{y} = (y_1, y_2, \ldots, y_N), \qquad y_i \in \mathbb{R} \quad \forall i$$

We wish to find a parametric function $f(x, \boldsymbol{\theta})$ that models the relationship between $\mathbf{x}$ and $\mathbf{y}$.

Linear regression: review

We assume that the data is generated according to the following model

$$y = f(x, \boldsymbol{\theta}) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2),$$

where σ^2 is a hyperparameter.

Therefore our assumption is that the data is generated by an underlying regression model, but observations are contaminated by Gaussian noise.

The random variable ϵ captures uncertainty due to inherent stochasticity of the process, measurement errors or just our ignorance of factors we are unable to capture.

Linear regression: review

We assume the following form for our regression function:

$$f(x, \boldsymbol{\theta}) = \sum_{m=1}^{M} \theta_m \phi_m(x)$$

where the basis functions $\phi_m(\cdot)$ define the features that we will use to make predictions.

The maximum likelihood estimator (MLE) for the parameters is given by

$$\begin{aligned} \boldsymbol{\theta}_{ML} &= \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) \quad \left(= \arg\max_{\boldsymbol{\theta}} p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) \right) \\ &= \arg\min_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} (y_i - \sum_{m=1}^{M} \theta_m \phi_m(x_i))^2 \end{aligned}$$

Linear regression: review

The MLE estimator is prone to overfitting, especially for small datasets. A typical Bayesian approach is to formulate the maximum a posteriori (MAP) estimator by first introducing a prior distribution over the model parameters:

$$\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}|\mathbf{0}, \alpha^2 \mathbf{I}_M)$$

The MAP estimator is then given by

$$\begin{aligned} \boldsymbol{\theta}_{MAP} &= \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) \\ &= \arg\min_{\boldsymbol{\theta}} \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - \sum_{m=1}^{M} \theta_m \phi_m(x_i))^2 + \frac{\lambda}{N} \sum_{m=1}^{M} \theta_m^2 \right\} \end{aligned}$$

where $\lambda = \sigma^2/\alpha^2$.

Linear regression: review

The MLE and MAP estimator can be computed in closed form by solving the (regularised) least squares problem:

$$\begin{aligned} \boldsymbol{\theta}_{ML} &= (\Phi_{\mathbf{x}}^T \Phi_{\mathbf{x}})^{-1} \Phi_{\mathbf{x}}^T \mathbf{y} \\ \boldsymbol{\theta}_{MAP} &= (\Phi_{\mathbf{x}}^T \Phi_{\mathbf{x}} + \lambda \mathbf{I}_M)^{-1} \Phi_{\mathbf{x}}^T \mathbf{y} \end{aligned}$$

where $\Phi_{\mathbf{x}} \in \mathbb{R}^{N \times M}$ is the design matrix, defined by

$$\Phi_{\mathbf{x}}[i, j] = \phi_j(x_i), \quad i = 1, \ldots, N, \, j = 1, \ldots, M.$$

Linear regression: review

The MLE or MAP estimators can be used to make model predictions. For a query input $x^* \in \mathbb{R}^d$, the model returns the prediction y^* given by

$$y^* = f(x^*, \boldsymbol{\theta}_{MAP}) = \boldsymbol{\phi}(x^*)^T \boldsymbol{\theta}_{MAP}$$

where $\boldsymbol{\phi}(x) = [\phi_1(x), \ldots, \phi_M(x)]^T$.

Recall that under the model this is in fact the prediction of the mean of the distribution $p(y|x^*, \boldsymbol{\theta}_{MAP})$:

$$\mathbb{E}[y|x^*, \boldsymbol{\theta}_{MAP}] = \boldsymbol{\phi}(x^*)^T \boldsymbol{\theta}_{MAP}$$

Linear regression: the Bayesian approach

A more thorough Bayesian treatment of the problem does not simply use a point estimate of the parameters $\boldsymbol{\theta}$, but instead accounts for the uncertainty in the true parameter values by integrating over all all possible values of $\boldsymbol{\theta}$.

The posterior distribution $p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y})$ reflects our belief about the likely parameter values, having seen the data.

The model prediction for a given value of $\boldsymbol{\theta}$ is given by $p(y|x, \boldsymbol{\theta})$.

Therefore the **predictive distribution** given a query input x^* is given by

$$p(y|x^*, \mathbf{x}, \mathbf{y}) = \int_{\boldsymbol{\theta}} p(y|x^*, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) d\boldsymbol{\theta}$$

We will derive the predictive distribution for our problem.

Outline

Multivariate Gaussian distribution

We first we recall the multivariate Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean $\boldsymbol{\mu} \in \mathbb{R}^n$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$. The probability density function is given by

$$p(\mathbf{x}) = (2\pi)^{-\frac{n}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

Joint Gaussian distribution

Let $\mathbf{x} \in \mathbb{R}^n$ be distributed according to a multivariate Gaussian distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^n$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$.

Suppose $\mathbf{x}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are partitioned as follows:

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix},$$

with $\mathbf{x}_1, \boldsymbol{\mu}_1 \in \mathbb{R}^q$, $\mathbf{x}_2, \boldsymbol{\mu}_2 \in \mathbb{R}^{n-q}$ and $\boldsymbol{\Sigma}_{11} \in \mathbb{R}^{q \times q}$, $\boldsymbol{\Sigma}_{12} \in \mathbb{R}^{q \times (n-q)}$, $\boldsymbol{\Sigma}_{21} \in \mathbb{R}^{(n-q) \times q}$ and $\boldsymbol{\Sigma}_{22} \in \mathbb{R}^{(n-q) \times (n-q)}$

We say the random variables $\mathbf{x}_1$ and $\mathbf{x}_2$ have a joint Gaussian distribution (or they are 'jointly Normally distributed').

Marginalisation

Suppose we would like to know the marginal distribution of $\mathbf{x}_1$:

$$p(\mathbf{x}_1) = \int_{\mathbb{R}^{(n-q)}} p(\mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}_2$$

$p(\mathbf{x}_1)$ is also a multivariate Gaussian density, with mean and covariance matrix equal to the corresponding elements of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$:

$$\mathbf{x}_1 \sim \mathcal{N}(\mathbf{x}_1 | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$$

Likewise for the marginal distribution of $\mathbf{x}_2$:

$$\mathbf{x}_2 \sim \mathcal{N}(\mathbf{x}_2 | \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$$

Conditional distributions

We may also be interested in the conditional distribution $p(\mathbf{x}_1|\mathbf{x}_2 = \mathbf{a})$. This is again a multivariate Gaussian distribution

$$p(\mathbf{x}_1|\mathbf{x}_2 = \mathbf{a}) = N(\mathbf{x}_1|\overline{\boldsymbol{\mu}}, \overline{\boldsymbol{\Sigma}})$$

where $N(\mathbf{x}, \boldsymbol{u}, \boldsymbol{\Sigma})$ denotes the multivariate Gaussian PDF. The mean $\overline{\boldsymbol{\mu}}$ and covariance matrix $\overline{\boldsymbol{\Sigma}}$ are given by

$$\begin{aligned} \overline{\boldsymbol{\mu}} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{a} - \boldsymbol{\mu}_2) \\ \overline{\boldsymbol{\Sigma}} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \end{aligned}$$

Note that the above is often just written as $p(\mathbf{x}_1|\mathbf{x}_2)$, and $\overline{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$.

Linear transformation of Gaussian random variables

Let $\mathbf{x} \in \mathbb{R}^n$ be a multivariate Gaussian random variable, $\mathbf{x} \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Let $\mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{b}$ be a linear transformation of $\mathbf{x}$ with $\mathbf{A} \in \mathbb{R}^{m \times n}$, $m \leq n$, $\mathbf{b} \in \mathbb{R}^m$ and assume $\mathbf{A}$ has full rank.

Then $\mathbf{z}$ is also a multivariate Gaussian random variable:

$$\mathbf{z} \sim \mathcal{N}(\mathbf{z}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$$

Note how the marginalisation property of the multivariate Gaussian is a special case of the above result.

Product of Gaussian densities

It will also turn out to be useful to calculate the product of Gaussian densities. In the following we denote the pdf as $N(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then we have

$$N(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)N(\mathbf{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) = \mathcal{Z}N(\mathbf{x}|\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$$

where

$$\tilde{\boldsymbol{\Sigma}} = (\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1})^{-1}$$

and

$$\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2)$$

The product of two Gaussian densities gives an unnormalised Gaussian density. The normalisation constant $\mathcal{Z}$ is given by

$$\mathcal{Z} = N(\boldsymbol{\mu}_1|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2) = N(\boldsymbol{\mu}_2|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$$

Outline

Posterior distribution $p(\theta|\mathbf{x}, \mathbf{y})$

Recall that for a fully Bayesian treatment of our linear regression problem, our aim is to find the predictive distribution given a query input x^*

$$p(y|x^*, \mathbf{x}, \mathbf{y}) = \int_{\boldsymbol{\theta}} p(y|x^*, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) d\boldsymbol{\theta}$$

which requires knowledge of the posterior distribution $p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y})$.

We have previously only derived point estimates for $\boldsymbol{\theta}$:

$$\begin{aligned} \boldsymbol{\theta}_{ML} &= \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) \quad (= p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})) \\ \boldsymbol{\theta}_{MAP} &= \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) \end{aligned}$$

To find the predictive distribution we will first derive the form of the posterior $p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y})$.

Posterior distribution

Recall that we have specified the prior distribution on our model parameters

$$\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}|\mathbf{0}, \alpha^2 \mathbf{I}_M)$$

and our model for the data is

$$y = f(x, \boldsymbol{\theta}) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

where $f(x, \boldsymbol{\theta}) = \sum_{m=1}^{M} \theta_m \phi_m(x) = \boldsymbol{\phi}(x)^T \boldsymbol{\theta}$.

Therefore given the dataset $(\mathbf{x}, \mathbf{y})$ we have

$$\mathbf{y} \sim \mathcal{N}(\mathbf{y}|\Phi_{\mathbf{x}}\boldsymbol{\theta}, \sigma^2 \mathbf{I}_N)$$

Posterior distribution

From Bayes' Theorem we have

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

The denominator $p(\mathbf{y}|\mathbf{x})$ is obtained by marginalising out $\boldsymbol{\theta}$ from the numerator. As it does not depend on $\boldsymbol{\theta}$ we will ignore it—we will compute the numerator $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})p(\boldsymbol{\theta})$ and then normalise the result to ensure it is a valid distribution.

Both $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$ are Gaussian probability density functions.

We can use our Gaussian identities to compute the numerator above!

Posterior distribution

First consider $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$. We write the Gaussian pdf as $N(\mathbf{y}|\Phi_{\mathbf{x}}\boldsymbol{\theta}, \sigma^2\mathbf{I}_N)$, so that

$$p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = N(\mathbf{y}|\Phi_{\mathbf{x}}\boldsymbol{\theta}, \sigma^2\mathbf{I}_N).$$

We assume the design matrix $\Phi_{\mathbf{x}}$ has full column rank (otherwise redundant features should be removed). In that case note that $\Phi_{\mathbf{x}}^T\Phi_{\mathbf{x}}$ is invertible (it is a Gram matrix, recall earlier discussion).

Now let $\mathbf{z} = (\Phi_{\mathbf{x}}^T\Phi_{\mathbf{x}})^{-1}\Phi_{\mathbf{x}}^T\mathbf{y}$. We use our earlier result on linear transformations of Gaussian random variables to see that

$$\mathbf{z} \sim \mathcal{N}(\mathbf{z}|(\Phi_{\mathbf{x}}^T\Phi_{\mathbf{x}})^{-1}\Phi_{\mathbf{x}}^T\Phi_{\mathbf{x}}\boldsymbol{\theta}, \left[(\Phi_{\mathbf{x}}^T\Phi_{\mathbf{x}})^{-1}\Phi_{\mathbf{x}}^T\right]\sigma^2\mathbf{I}_N\left[(\Phi_{\mathbf{x}}^T\Phi_{\mathbf{x}})^{-1}\Phi_{\mathbf{x}}^T\right)^T])$$

$$\Rightarrow \mathbf{z} \sim \mathcal{N}(\mathbf{z}|\boldsymbol{\theta}, \sigma^2(\Phi_{\mathbf{x}}^T\Phi_{\mathbf{x}})^{-1})$$

Posterior distribution

As $\mathbf{z}$ is a deterministic (linear) transformation of $\mathbf{y}$, it is clear that

$$\begin{aligned} p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta}) &= p(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}) \\ &= N(\mathbf{z}|\boldsymbol{\theta},\sigma^2(\Phi_{\mathbf{x}}^T\Phi_{\mathbf{x}})^{-1}) \\ &= N((\Phi_{\mathbf{x}}^T\Phi_{\mathbf{x}})^{-1}\Phi_{\mathbf{x}}^T\mathbf{y}|\boldsymbol{\theta},\sigma^2(\Phi_{\mathbf{x}}^T\Phi_{\mathbf{x}})^{-1}) \\ &= N(\boldsymbol{\theta}|(\Phi_{\mathbf{x}}^T\Phi_{\mathbf{x}})^{-1}\Phi_{\mathbf{x}}^T\mathbf{y},\sigma^2(\Phi_{\mathbf{x}}^T\Phi_{\mathbf{x}})^{-1}) \end{aligned}$$

and therefore

$$p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})p(\boldsymbol{\theta}) = N(\boldsymbol{\theta}|(\Phi_{\mathbf{x}}^T\Phi_{\mathbf{x}})^{-1}\Phi_{\mathbf{x}}^T\mathbf{y},\sigma^2(\Phi_{\mathbf{x}}^T\Phi_{\mathbf{x}})^{-1})N(\boldsymbol{\theta}|\mathbf{0},\alpha^2\mathbf{I}_M)$$

Posterior distribution

$$p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})p(\boldsymbol{\theta}) = N(\boldsymbol{\theta}|(\boldsymbol{\Phi}_{\mathbf{x}}^T\boldsymbol{\Phi}_{\mathbf{x}})^{-1}\boldsymbol{\Phi}_{\mathbf{x}}^T\mathbf{y}, \sigma^2(\boldsymbol{\Phi}_{\mathbf{x}}^T\boldsymbol{\Phi}_{\mathbf{x}})^{-1})N(\boldsymbol{\theta}|\mathbf{0}, \alpha^2\mathbf{I}_M)$$

Note the above is in the form where we are able to invoke our result about the product of two Gaussian densities. Therefore we obtain

$$p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})p(\boldsymbol{\theta}) = \mathcal{Z}N(\boldsymbol{\theta}|\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$$

where $\mathcal{Z}$ is a constant (independent of $\boldsymbol{\theta}$), and

$$\begin{aligned}
\tilde{\boldsymbol{\Sigma}} &= (\sigma^{-2}(\boldsymbol{\Phi}_{\mathbf{x}}^T\boldsymbol{\Phi}_{\mathbf{x}}) + \alpha^{-2}\mathbf{I}_M)^{-1} \\
&= \sigma^2(\boldsymbol{\Phi}_{\mathbf{x}}^T\boldsymbol{\Phi}_{\mathbf{x}} + (\sigma^2/\alpha^2)\mathbf{I}_M)^{-1} \\
\tilde{\boldsymbol{\mu}} &= \tilde{\boldsymbol{\Sigma}}(\sigma^{-2}(\boldsymbol{\Phi}_{\mathbf{x}}^T\boldsymbol{\Phi}_{\mathbf{x}})(\boldsymbol{\Phi}_{\mathbf{x}}^T\boldsymbol{\Phi}_{\mathbf{x}})^{-1}\boldsymbol{\Phi}_{\mathbf{x}}^T\mathbf{y}) \\
&= \tilde{\boldsymbol{\Sigma}}(\sigma^{-2}\boldsymbol{\Phi}_{\mathbf{x}}^T\mathbf{y}) \\
&= (\boldsymbol{\Phi}_{\mathbf{x}}^T\boldsymbol{\Phi}_{\mathbf{x}} + (\sigma^2/\alpha^2)\mathbf{I}_M)^{-1}\boldsymbol{\Phi}_{\mathbf{x}}^T\mathbf{y}
\end{aligned}$$

Posterior distribution

As before, we set $\lambda = \sigma^2/\alpha^2$, so we have

$$p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})p(\boldsymbol{\theta}) = \mathcal{Z}N(\boldsymbol{\theta}|(\Phi_{\mathbf{x}}^T\Phi_{\mathbf{x}} + \lambda\mathbf{I}_M)^{-1}\Phi_{\mathbf{x}}^T\mathbf{y}, \sigma^2(\Phi_{\mathbf{x}}^T\Phi_{\mathbf{x}} + \lambda\mathbf{I}_M)^{-1})$$

Now recall that

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

and so we can see from the above that the posterior $p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y})$ is also a multivariate Gaussian:

$$p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) \sim \mathcal{N}(\boldsymbol{\theta}|(\Phi_{\mathbf{x}}^T\Phi_{\mathbf{x}} + \lambda\mathbf{I}_M)^{-1}\Phi_{\mathbf{x}}^T\mathbf{y}, \sigma^2(\Phi_{\mathbf{x}}^T\Phi_{\mathbf{x}} + \lambda\mathbf{I}_M)^{-1})$$

We recognise the mean of the distribution as the MAP estimate for $\boldsymbol{\theta}$ (a.k.a. regularised least squares solution).

Predictive distribution

Now we have computed the posterior distribution $p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y})$, we would like to evaluate the predictive distribution

$$p(y|x^*, \mathbf{x}, \mathbf{y}) = \int p(y|x^*, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) d\boldsymbol{\theta}$$

To compute the above distribution we marginalise out the parameters $\boldsymbol{\theta}$. We will do this by computing the joint distribution for $(y, \boldsymbol{\theta})$ and then using our earlier Gaussian identity of marginalisation of variables.

Predictive distribution: general multivariate Gaussian

We begin by expanding the probability density function for a general multivariate Gaussian variable $\mathbf{z} \sim \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$\begin{aligned} p(\mathbf{z}) &= (2\pi)^{-\frac{n}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})\right) \\ &= C_1 \exp\left(-\frac{1}{2}\mathbf{z}\boldsymbol{\Sigma}^{-1}\mathbf{z} + \mathbf{z}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\mu}\right) \\ &= C_2 \exp\left(-\frac{1}{2}\mathbf{z}\boldsymbol{\Sigma}^{-1}\mathbf{z} + \mathbf{z}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \end{aligned}$$

where the constants C_1, C_2 are independent of $\mathbf{z}$. In the above we have used the fact that $\boldsymbol{\Sigma}$ (and therefore $\boldsymbol{\Sigma}^{-1}$) is symmetric.

Predictive distribution: joint distribution of (θ, y)

We now compute the joint distribution

$$\begin{aligned} p(\boldsymbol{\theta}, y|x^*, \mathbf{x}, \mathbf{y}) &= p(y|x^*, \boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) \\ &= K_1 \exp\left\{-\frac{1}{2}\sigma^{-2}(y - \boldsymbol{\phi}(x^*)^T\boldsymbol{\theta})^2 \right. \\ &\qquad \left. -\frac{1}{2}\left(\boldsymbol{\theta} - \mathbf{A}^{-1}\Phi_{\mathbf{x}}^T\mathbf{y}\right)^T \sigma^{-2}\mathbf{A}\left(\boldsymbol{\theta} - \mathbf{A}^{-1}\Phi_{\mathbf{x}}^T\mathbf{y}\right)\right\} \end{aligned}$$

where $\mathbf{A} = \Phi_{\mathbf{x}}^T\Phi_{\mathbf{x}} + \lambda\mathbf{I}_M$. and the constant is independent of both y and $\boldsymbol{\theta}$.

Predictive distribution: joint distribution of (θ, y)

Now we collect together quadratic and linear terms, and absorb any terms independent of both y and $\boldsymbol{\theta}$ into the constant:

$$p(\boldsymbol{\theta}, y|x^*, \mathbf{x}, \mathbf{y}) = K_2 \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{\theta}^T(\mathbf{A}+\boldsymbol{\phi}\boldsymbol{\phi}^T)\boldsymbol{\theta} - 2y\boldsymbol{\phi}(x^*)^T\boldsymbol{\theta} + y^2 - 2\boldsymbol{\theta}^T\Phi_{\mathbf{x}}^T\mathbf{y}\right)\right\}$$

$$= K_2 \exp\left\{-\frac{1}{2\sigma^2}\begin{bmatrix}\boldsymbol{\theta}\\ y\end{bmatrix}^T\begin{bmatrix}\mathbf{A}+\boldsymbol{\phi}(x^*)\boldsymbol{\phi}(x^*)^T & -\boldsymbol{\phi}(x^*)\\ -\boldsymbol{\phi}(x^*)^T & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{\theta}\\ y\end{bmatrix} + \frac{1}{\sigma^2}\begin{bmatrix}\boldsymbol{\theta}\\ y\end{bmatrix}^T\begin{bmatrix}\Phi_{\mathbf{x}}^T\mathbf{y}\\ 0\end{bmatrix}\right\}$$

and we can see that the joint distribution of $(\boldsymbol{\theta}, y)$ is also multivariate Gaussian.

Predictive distribution: covariance matrix

By comparing quadratic terms to the expression for a general multivariate Gaussian variable $\mathbf{z}$, we see that the covariance matrix (inverse of the **precision matrix**) of the (θ, y) joint distribution is equal to

$$\sigma^2 \begin{bmatrix} \mathbf{A} + \phi(x^*)\phi(x^*)^T & -\phi(x^*) \\ -\phi(x^*)^T & 1 \end{bmatrix}^{-1}$$

This inverse can be computed using a blockwise inversion identity, and we find the covariance matrix is equal to:

$$\begin{bmatrix} \sigma^2\mathbf{A}^{-1} & \sigma^2\mathbf{A}^{-1}\phi(x^*) \\ \sigma^2\phi(x^*)^T\mathbf{A}^{-1} & \sigma^2 + \sigma^2\phi(x^*)^T\mathbf{A}^{-1}\phi(x^*) \end{bmatrix}$$

Predictive distribution: mean

Likewise, by comparing linear terms we see that the mean of the (θ, y) distribution is given by

$$\begin{bmatrix} \mathbf{A}^{-1} & \mathbf{A}^{-1}\phi(x^*) \\ \phi(x^*)^T\mathbf{A}^{-1} & 1 + \phi(x^*)^T\mathbf{A}^{-1}\phi(x^*) \end{bmatrix} \begin{bmatrix} \Phi_{\mathbf{x}}^T\mathbf{y} \\ 0 \end{bmatrix}$$
$$= \begin{bmatrix} \mathbf{A}^{-1}\Phi_{\mathbf{x}}^T\mathbf{y} \\ \phi(x^*)^T\mathbf{A}^{-1}\Phi_{\mathbf{x}}^T\mathbf{y} \end{bmatrix}$$

Predictive distribution: marginal distribution

Finally, we can now write down the marginal distribution $p(y|x^*, \mathbf{x}, \mathbf{y})$ using the earlier identity that the marginal is simply given by the corresponding mean vector and covariance block. Thus we have:

$$p(y|x^*, \mathbf{x}, \mathbf{y}) \sim \mathcal{N}(y|\overline{\mu}, \overline{\sigma}^2)$$

where the mean is given by

$$\overline{\mu} = \phi(x^*)^T(\Phi_{\mathbf{x}}^T\Phi_{\mathbf{x}} + \lambda \mathbf{I}_M)^{-1}\Phi_{\mathbf{x}}^T\mathbf{y}$$

and the variance (note $y \in \mathbb{R}$) is

$$\overline{\sigma}^2 = \sigma^2 + \sigma^2\phi(x^*)^T(\Phi_{\mathbf{x}}^T\Phi_{\mathbf{x}} + \lambda \mathbf{I}_M)^{-1}\phi(x^*))$$

Predictive distribution: mean and variance

We can see that the mean of the predictive distribution is equal to our model prediction using the MAP estimate:

$$\phi(x^*)^T(\Phi_{\mathbf{x}}^T\Phi_{\mathbf{x}} + \lambda \mathbf{I}_M)^{-1}\Phi_{\mathbf{x}}^T\mathbf{y} = \phi(x^*)^T\boldsymbol{\theta}_{MAP} = f(x^*, \boldsymbol{\theta}_{MAP})$$

We also notice that the variance of the predictive distribution splits into two parts: the first term comes from the irreducible noise that we assumed as part of our model. The second term comes from uncertainty about the model parameters $\boldsymbol{\theta}$:

$$\underbrace{\sigma^2}_{\text{data uncertainty}} + \underbrace{\sigma^2\phi(x^*)^T(\Phi_{\mathbf{x}}^T\Phi_{\mathbf{x}} + \lambda \mathbf{I}_M)^{-1}\phi(x^*))}_{\text{model uncertainty}}$$

Predictive distribution: mean and variance

- The two sources of variance (data noise process and model parameter uncertainty) are independent Gaussians, which is why the variances are additive
- The first term is irreducible; no amount of data or knowledge about $\boldsymbol{\theta}$ will change the noise in the observations
- As more data is observed, we become less uncertain about $\boldsymbol{\theta}$ and the posterior variance narrows
- As a result, it can be shown that the predictive distribution variance also narrows with more data. In the limit $N \to \infty$, the second term vanishes and the variance consists solely of the observation noise

Outline

Summing it all up

Let us once again recap what we have done so far. We assumed a model of the form

$$y = f(x, \boldsymbol{\theta}) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

to fit our dataset $(\mathbf{x}, \mathbf{y})$, where $f(x, \boldsymbol{\theta}) = \sum_{m=1}^{M} \theta_m \phi_m(x)$ and $\phi_m(\cdot)$ are a set of (in general nonlinear) basis functions.

We derived two point estimates for $\boldsymbol{\theta}$. The maximum likelihood estimate (MLE) given by

$$\boldsymbol{\theta}_{ML} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) \quad \left(= \arg\max_{\boldsymbol{\theta}} p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) \right)$$

and the maximum a posteriori (MAP) estimate, given by

$$\boldsymbol{\theta}_{MAP} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y})$$

Summing it all up

For either of these point estimates, the predictive distribution $p(y|x^*, \boldsymbol{\theta})$ for a query input is given by

$$p(y|x^*, \boldsymbol{\theta}) \sim \mathcal{N}(y|\boldsymbol{\phi}(x^*)^T\boldsymbol{\theta}, \sigma^2)$$

where $\boldsymbol{\phi}(x) = [\phi_1(x), \ldots, \phi_M(x)]^T$ and $\boldsymbol{\theta} = \boldsymbol{\theta}(\mathbf{x}, \mathbf{y})$.

The variance in the predictive distribution is purely down to intrinsic noise in the observations (e.g. due to inherent stochasticity, measurement noise or other factors of variation).

The idea behind Bayesian linear regression is to account for uncertainty in the model parameters $\boldsymbol{\theta}$.

Summing it all up

We place a prior distribution over $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}|\mathbf{0}, \alpha^2 \mathbf{I}_M)$$

and compute the posterior distribution over the model parameters:

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

Given the posterior distribution, our predictive distribution now accounts for the uncertainty in $\boldsymbol{\theta}$ by integrating over the model parameters:

$$p(y|x^*, \mathbf{x}, \mathbf{y}) = \int p(y|x^*, \boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y})d\boldsymbol{\theta}$$

Note how the predictive distribution no longer depends on $\boldsymbol{\theta}$!

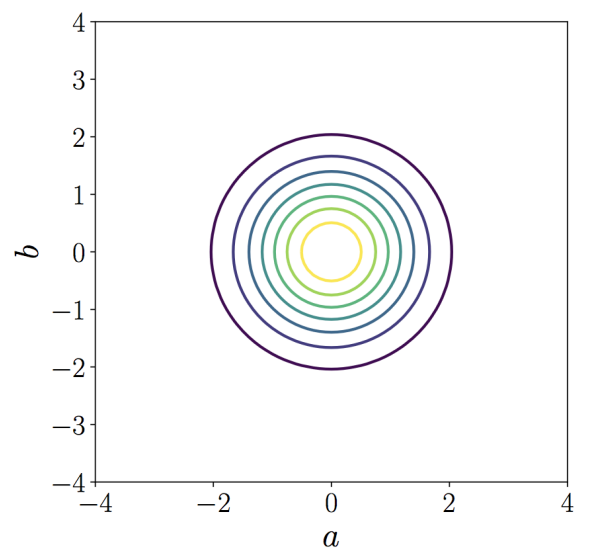
Outline

Example 1: 2D parameter space

We will illustrate the concepts in this lecture with a simple linear regression example. Our model is given by

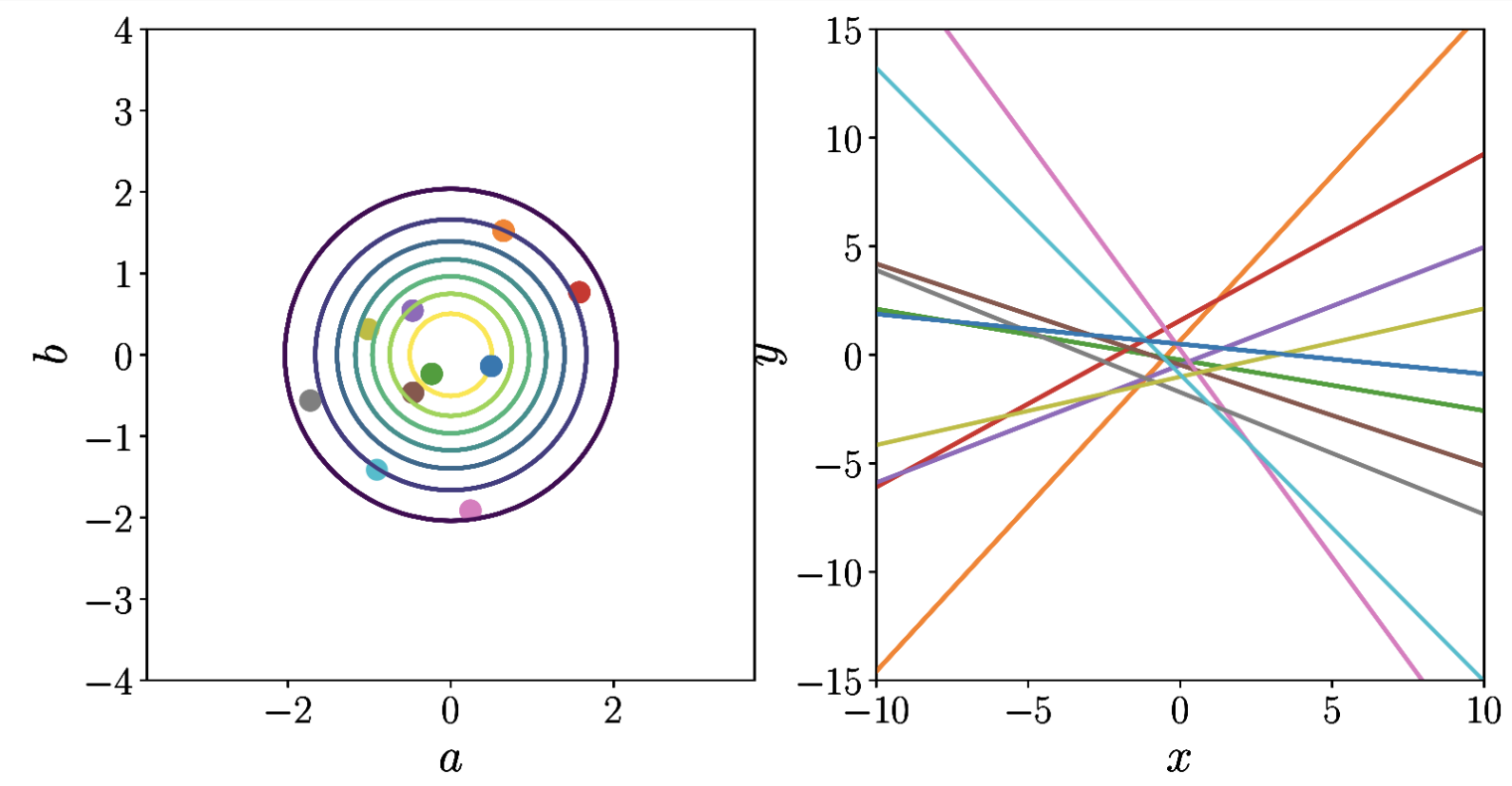
$$\begin{aligned} y &= f(x, \boldsymbol{\theta}) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2) \\ &= a + bx + \epsilon \end{aligned}$$

with our parameters $a, b \in \mathbb{R}$, our input $x \in \mathbb{R}$ and we specify the prior on our parameters as $[a, b]^T \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$.



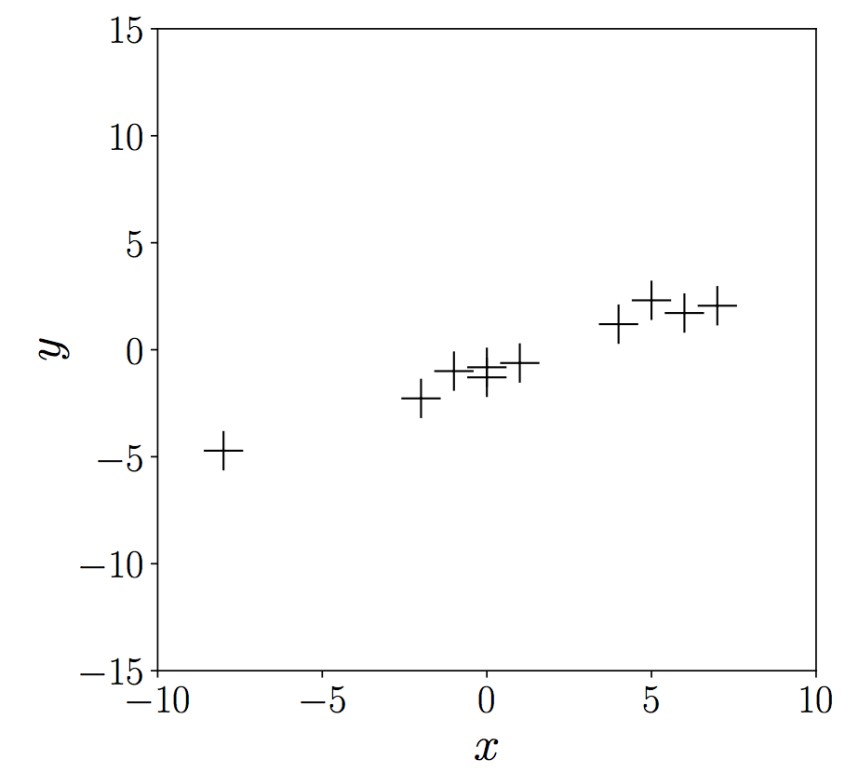
Example 1: sampling from the prior

We can sample parameters from our prior, which amounts to sampling straight line models:



Example 1: data

Now we suppose we are given a dataset consisting of example inputs $\mathbf{x} = [x_1, \ldots, x_N]^T$ and outputs $\mathbf{y} = [y_1, \ldots, y_N]^T$:



Example 1: computing the posterior

We now compute the posterior distribution $p(a, b|\mathbf{x}, \mathbf{y})$ by first assembling the design matrix

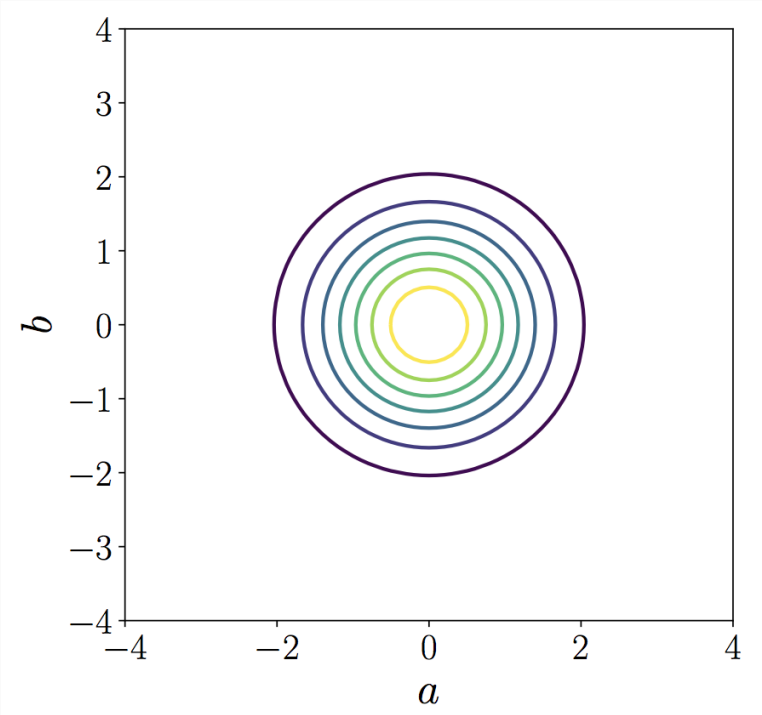
$$\Phi_{\mathbf{x}} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}$$

and setting $\lambda = \sigma^2$, we compute the posterior distribution

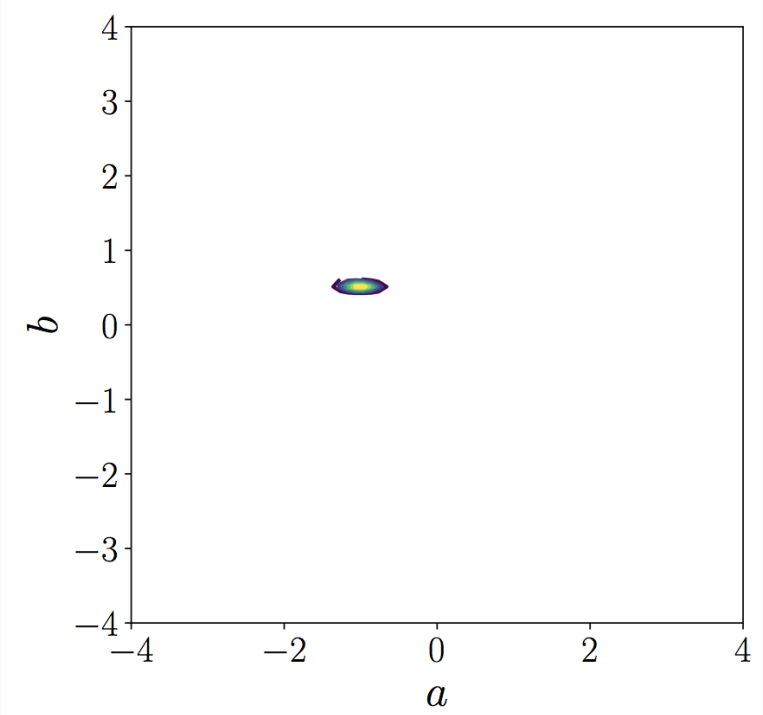
$$p([a, b]^T|\mathbf{x}, \mathbf{y}) \sim \mathcal{N}([a, b]^T|(\Phi_{\mathbf{x}}^T\Phi_{\mathbf{x}} + \lambda\mathbf{I}_M)^{-1}\Phi_{\mathbf{x}}^T\mathbf{y}, \sigma^2(\Phi_{\mathbf{x}}^T\Phi_{\mathbf{x}} + \lambda\mathbf{I}_M)^{-1})$$

Example 1: visualising the posterior

The posterior distribution over $\boldsymbol{\theta} = [a, b]^T$ represents our updated belief about the model parameters, given the data:



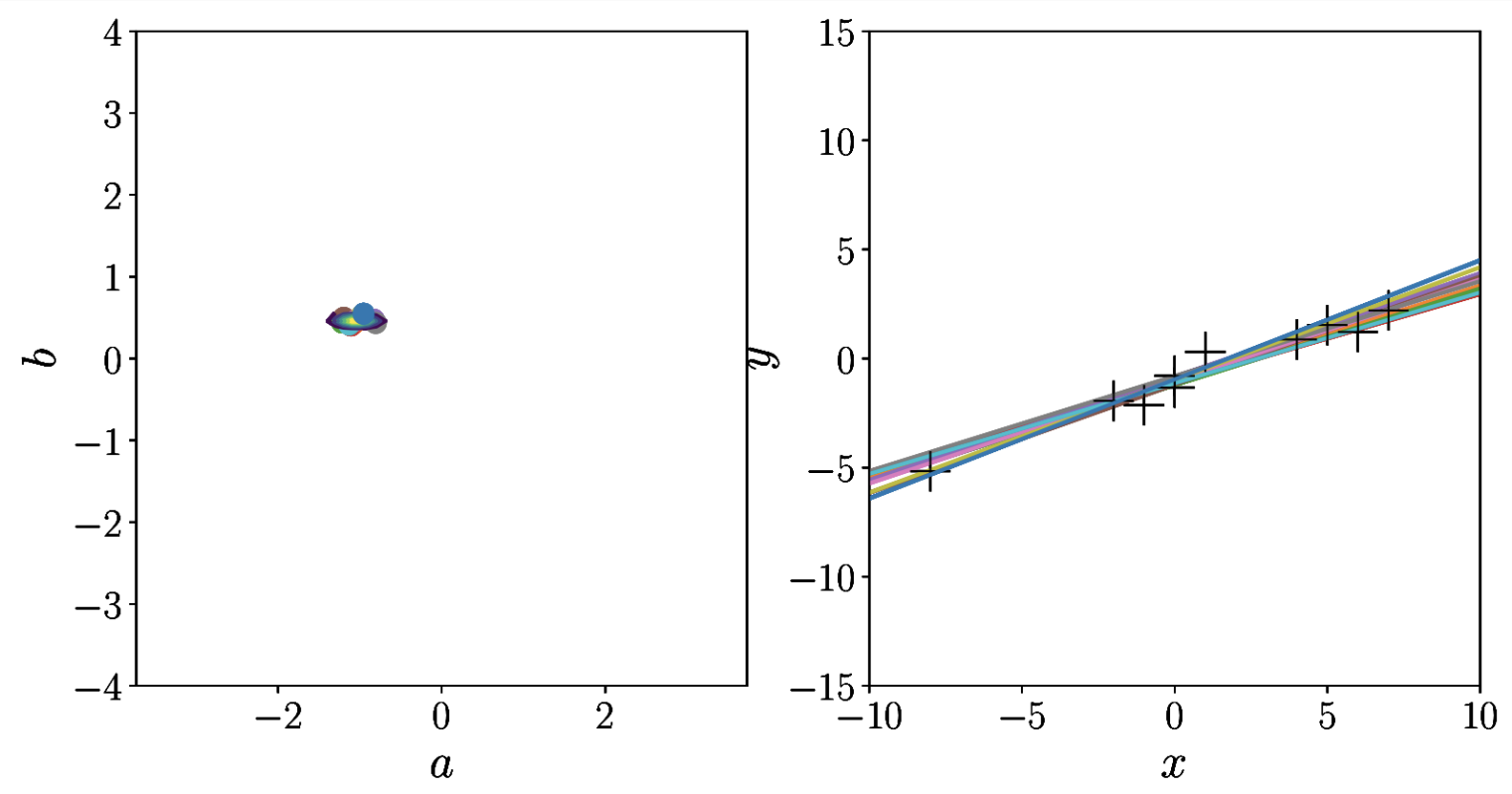
Prior



Posterior

Example 1: sampling from the posterior

We can now sample parameters from the posterior distribution:



Example 2: nonlinear model

A nonlinear model (in the inputs) is similar—in this case, our model takes the form

$$f(x, \boldsymbol{\theta}) = \sum_{m=1}^{M} \theta_m \phi_m(x), \qquad x \in \mathbb{R}$$

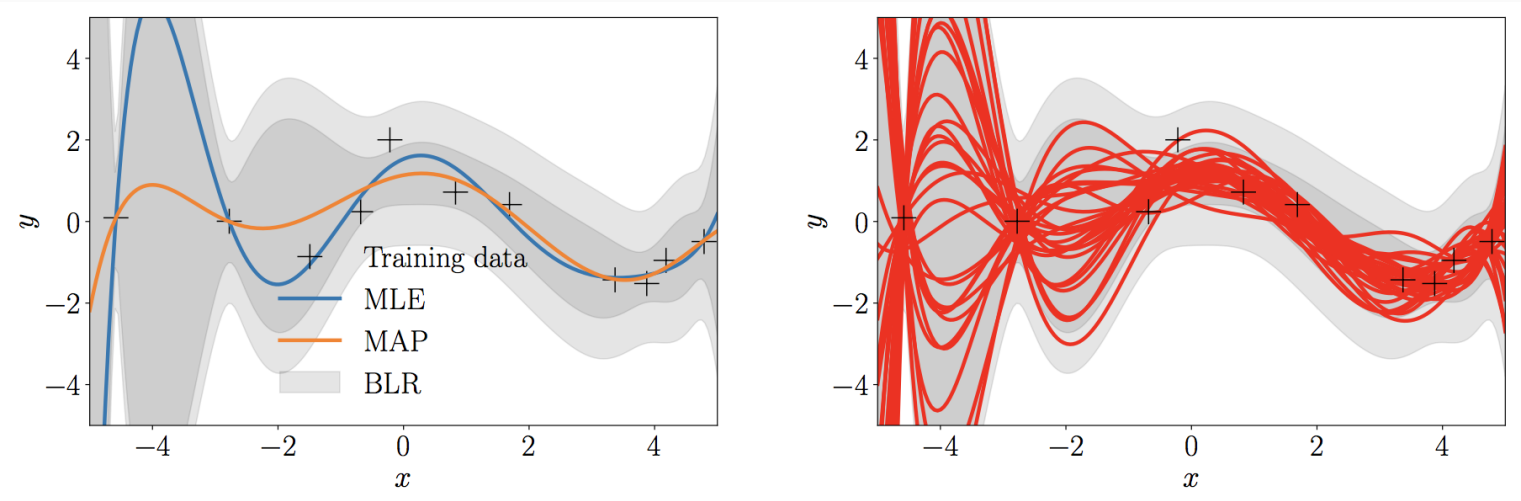
where $\phi_m(\cdot)$ are nonlinear basis functions (e.g. RBF functions $\phi_m(x) = \exp(-\frac{1}{2}(x - \mu_m)^2)$). Then the design matrix takes the form

$$\Phi_{\mathbf{x}} = \begin{bmatrix} \phi_1(x_1) & \cdots & \phi_M(x_1) \\ \vdots & \ddots & \vdots \\ \phi_1(x_N) & \cdots & \phi_M(x_N) \end{bmatrix}$$

and we calculate the posterior distribution in the same way.

Example 2: predictive distribution

The predictive distribution provides a mean and variance for a given input:



Note the MAP solution is the mean of the (fully Bayesian) predictive distribution, and is more regularised than the MLE solution.

Light grey indicates (irreducible) data uncertainty, dark grey indicates model uncertainty. The right hand plot shows functions sampled from the posterior.