Lecture 1b: Linear Models for Regression

Cédric Archambeau

Centre for Computational Statistics and Machine Learning
Department of Computer Science
University College London

c.archambeau@cs.ucl.ac.uk

Advanced Topics in Machine Learning (MSc in Intelligent Systems)
January 2008

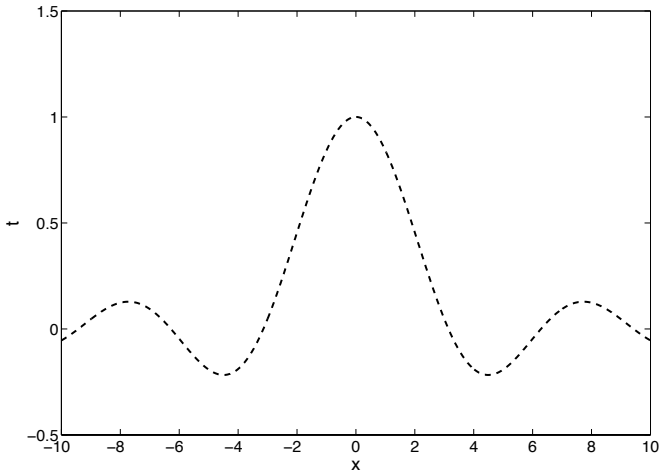
Today's plan

- Least squares and ridge regression: a probabilistic view
- Bayesian linear models for regression
- Sparse extensions
- Maximum likelihood, maximum a posteriori, type II ML, variational inference

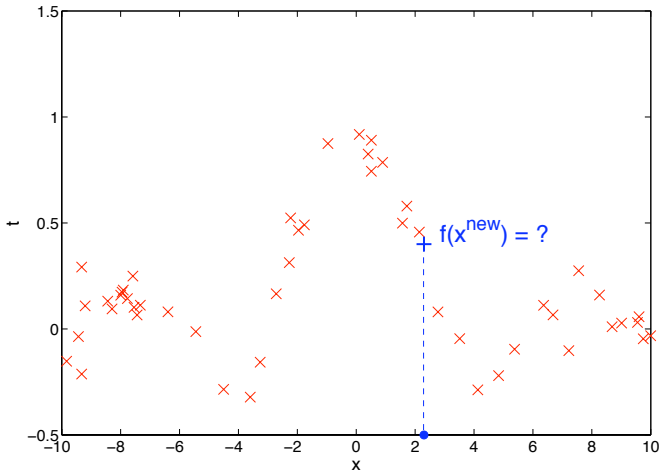
Regression problem

Given a **finite** number of **noisy** observations $\{t_n\}_{n=1}^N$ associated to some input data $\{\mathbf{x}_n\}_{n=1}^N$, we would like to predict the outcome of an unseen input $\mathbf{x}^{\text{new}}$.

This process is called **generalisation** and the input-target pairs $\{\mathbf{x}_n, t_n\}_{n=1}^N$ are the **training data**.



(a) Target.



(b) Generalisation.

Linear models for regression

The model $y(\mathbf{x}, \mathbf{w})$ is linear in the parameters $\mathbf{w}$ and is expressed as a weighted sum of nonlinear basis functions $\{\phi_m(\cdot)\}_{m=1}^{M}$ centred on M learning prototypes:

$$y(\mathbf{x}, \mathbf{w}) = \sum_{m=1}^{M} w_m \phi_m(\mathbf{x}) + w_0 = \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}).$$

- Least squares regression
- Partial least squares
- Regularization networks
- Support vector machines
- Radial basis function networks
- Splines
- Lasso
- ...

The goal is to **infer the parameters** from a set of real valued input-target pairs $\{\mathbf{x}_n, t_n\}_{n=1}^{N}$ which lead to the best prediction on unseen data.

Least squares regression

We consider the sum-of-squares error:

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(t_n - y_n)^2 = \frac{1}{2}\|\mathbf{t} - \mathbf{y}\|^2,$$

where $y_n \equiv y(\mathbf{x}_n, \mathbf{w})$, $\mathbf{t} = (t_1, \ldots, t_N)^\top$ and $\mathbf{y} = (y_1, \ldots, y_N)^\top$.

Since this expression is quadratic in $\mathbf{w}$, its minimisation leads to a unique minimum:

$$\mathbf{w}_{\mathrm{LS}} = (\boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^\top \mathbf{t} = \boldsymbol{\Phi}^\dagger \mathbf{t},$$

where $\boldsymbol{\Phi}^\dagger$ is the *Moore-Penrose pseudo-inverse* of $\boldsymbol{\Phi} \in \mathbb{R}^{N \times (M+1)}$.

In practice, $\boldsymbol{\Phi}$ is often ill-conditioned and solving the linear system leads to **overfitting** (low bias, but high variance; too much flexibility!).

The design matrix $\mathbf{\Phi}$ is defined as follows:

$$\mathbf{\Phi} = \begin{pmatrix} 1 & \phi_1(\mathbf{x}_1) & \dots & \phi_M(\mathbf{x}_1) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \phi_1(\mathbf{x}_N) & \dots & \phi_M(\mathbf{x}_N) \end{pmatrix}.$$

Using this definition we can write $\mathbf{y} = \mathbf{\Phi}\mathbf{w}$.

Standard **linear regression** is recovered for $\mathbf{\Phi} = \mathbf{X}$.

Example of overfitting

The target function is given by

$$f(x) = \frac{\sin x}{x}, \quad x \in [-10, 10]\,.$$

We choose the **squared exponential** basis function:

$$\phi_m(x) = \exp\left\{-\frac{\lambda_m}{2}(x - x_m)^2\right\}, \quad \lambda_m > 0.$$

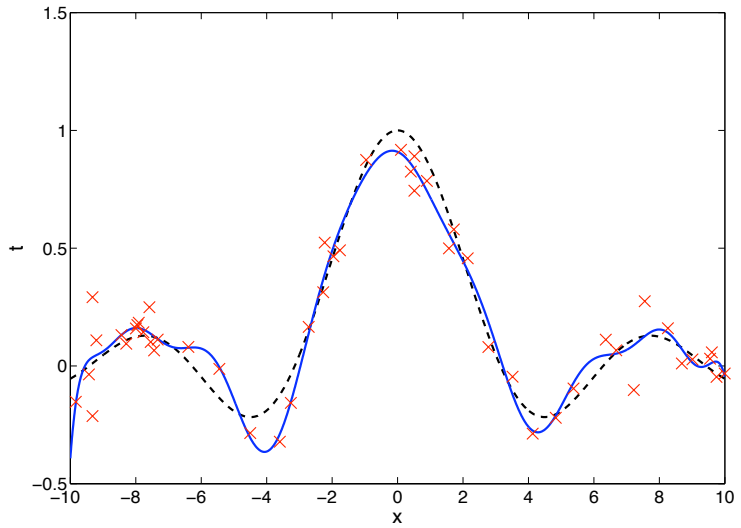


Figure: The target sinc function (dashed line) and the least squares regression solution (solid line) for $\lambda_m = 1/36$ for all m. The noisy observations are denoted by crosses.

Ridge regression (or weight decay)

The idea is to introduce **regularisation**, i.e. favour smooth regressors by penalising large $\|\mathbf{w}\|$.

The error function for ridge regression is defined as follows:

$$\widetilde{E}(\mathbf{w}) = E(\mathbf{w}) + \frac{\alpha}{2}\|\mathbf{w}\|^2,$$

where $\alpha \geqslant 0$ is the complexity parameter.

The global optimum for $\mathbf{w}$ is given by

$$\mathbf{w}_{\mathrm{R}} = (\mathbf{\Phi}^\top \mathbf{\Phi} + \alpha \mathbf{I}_M)^{-1} \mathbf{\Phi}^\top \mathbf{t}.$$

Hence, α converts an ill-conditioned problem into a well-conditioned one.

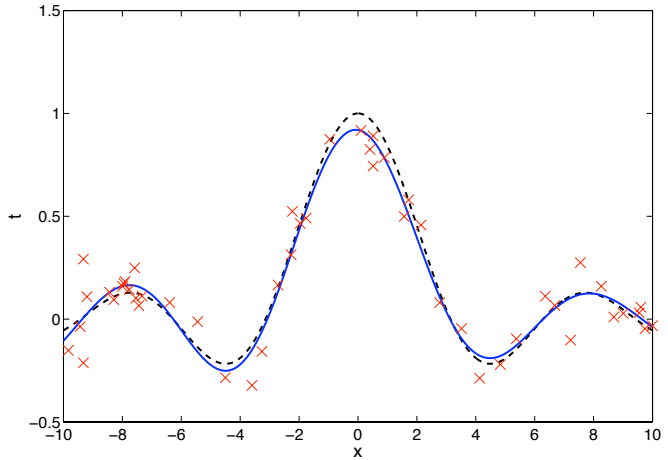
The effective **complexity** of the model is **reduced**, avoiding overfitting.

Example revisited

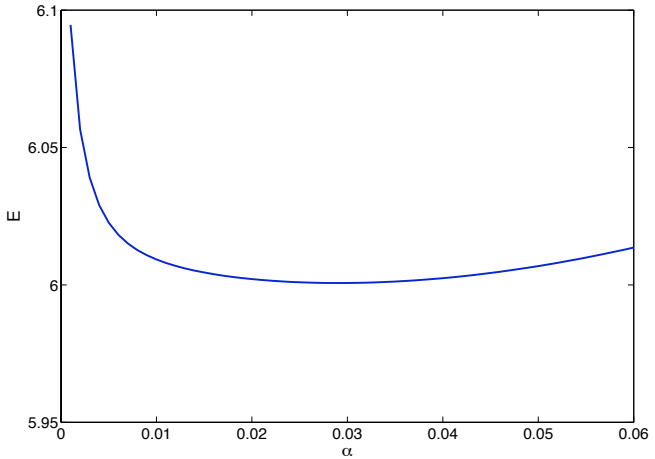
The target function is still given by

$$f(x) = \frac{\sin x}{x}, \quad x \in [-10, 10]\,.$$

The amount of penalisation depends on the value of α.



(a)



(b)

Figure: (a) The target sinc function (dashed line) and the least squares regression solution (solid line) for $\lambda_m = 1/36$ for all m. The noisy observations are denoted by crosses. (b) Penalised error as a function of α.

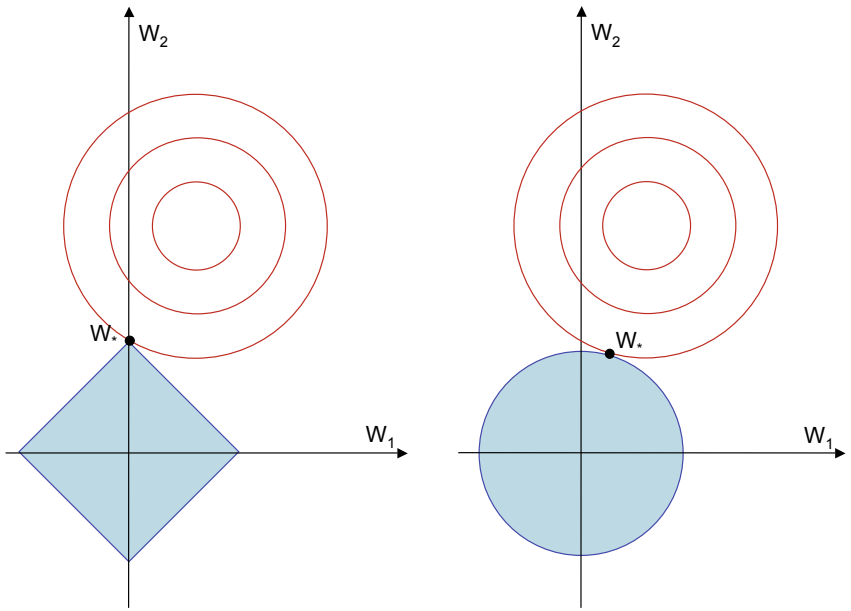
Parameter shrinkage and the LASSO

More general penalised error functions are of the following form:

$$\underset{\mathbf{w}}{\operatorname{argmin}}\ E(\mathbf{w}) \quad \text{subject to} \quad \sum_{m=0}^{M} |w_m|^q \leqslant \eta,$$

where q defines the type of regulariser:

- Ridge regression corresponds to the specific choice $q = 2$.
- The LASSO corresponds to $q = 1$ and leads to **sparse solutions** for a sufficiently small η as most weights are driven to zero.



Why using a probabilistic formalism in regression problems?

- Well-known least squares and ridge regression are special cases.
- The approach provides additional insights in the solution by dealing with the **uncertainty** in a principled way.
- Various sources of uncertainty can be modelled.
- **Confidence measures** are associated to the predictions that are made.
- Intensive resampling techniques such as cross-validation and the bootstrap are avoided.

What do we loose?

In practice, probabilistic and Bayesian inference require the computation of **intractable integrals** (marginalisation and normalisation/partition function).

These integrals are either estimated by *Markov Chain Monte Carlo* (MCMC) or by an *approximate inference* procedure (e.g. Laplace approximation, variational Bayes, expectation-propagation).

A probabilistic view of least squares regression

We assume that the observations are noisy iid samples drawn from a (univariate) Gaussian:

$$t_n = y_n + \epsilon_n, \quad \epsilon_n \sim \mathcal{N}(0, \sigma^2).$$

The likelihood of the observations is then given by a multivariate Gaussian:

$$(t_1, \ldots, t_N)|\mathbf{w}, \sigma \sim \prod_{n=1}^{N} \mathcal{N}(y_n, \sigma^2) = \mathcal{N}(\mathbf{y}, \sigma^2 \mathbf{I}_N).$$

The **maximum likelihood** (ML) solution leads to a set of equations:

$$\frac{d \ln p(\mathbf{t}|\mathbf{w}, \sigma)}{d\mathbf{w}} = 0 \quad \Rightarrow \quad \mathbf{w}_{\mathrm{ML}} = \mathbf{\Phi}^{\dagger}\mathbf{t},$$

$$\frac{d \ln p(\mathbf{t}|\mathbf{w}, \sigma)}{d\sigma^2} = 0 \quad \Rightarrow \quad \sigma^2_{\mathrm{ML}} = \frac{1}{N}\|\mathbf{t} - \mathbf{y}\|^2.$$

The ML solution for $\mathbf{w}$ is equal to the least squares solution. The ML estimate for the noise variance is to the residual error or **unexplained variance**.

The log-likelihood is given by

$$\ln p(\mathbf{t}|\mathbf{w},\sigma) = -\frac{N}{2}\ln 2\pi - \frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\underbrace{(\mathbf{t}-\mathbf{y})^\top(\mathbf{t}-\mathbf{y})}_{=\|\mathbf{t}-\mathbf{y}\|^2}.$$

Hence, this leads to

$$\frac{d\ln p(\mathbf{t}|\mathbf{w},\sigma)}{d\mathbf{w}} = 0 \quad \Rightarrow \quad \frac{1}{\sigma^2}\boldsymbol{\Phi}^\top(\mathbf{t}-\boldsymbol{\Phi}\mathbf{w}) = 0,$$

$$\frac{d\ln p(\mathbf{t}|\mathbf{w},\sigma)}{d\sigma^2} = 0 \quad \Rightarrow \quad -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{t}-\mathbf{y})^\top(\mathbf{t}-\mathbf{y}) = 0.$$

A probabilistic view of ridge regression

How to avoid overfitting?

We model the uncertainty on the value of the parameters by imposing some prior distribution on them:

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1}),$$

where $\mathbf{A} \equiv \mathrm{diag}\{\alpha_0, \ldots, \alpha_M\}$.

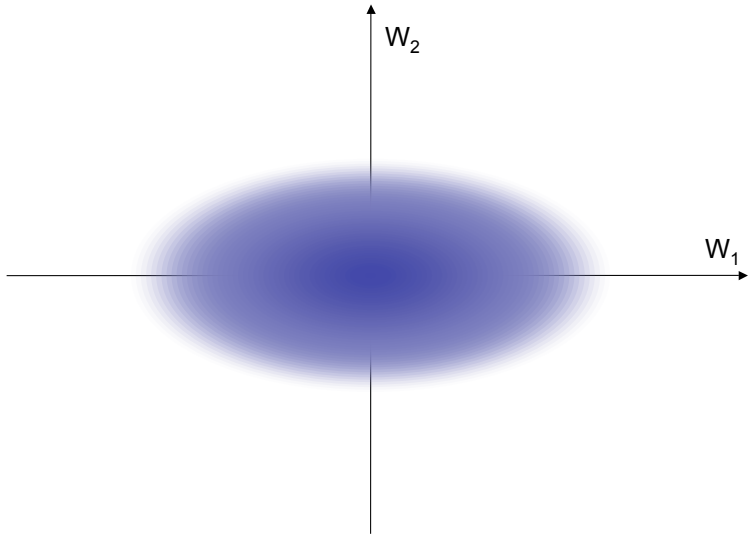


Figure: Zero-mean Gaussian prior with diagonal covariance matrix.

A probabilistic view of ridge regression (continued)

Applying *Bayes' rule* leads to the **posterior** distribution of the parameters:

$$p(\mathbf{w}|\mathbf{t}) \propto p(\mathbf{t}|\mathbf{w})p(\mathbf{w}).$$

The **maximum a posteriori** (MAP) solution is given by

$$\frac{d \ln p(\mathbf{w}|\mathbf{t})}{d\mathbf{w}} = 0 \quad \Rightarrow \quad \mathbf{w}_{\mathrm{MAP}} = \sigma^{-2}(\sigma^{-2}\mathbf{\Phi}^\top \mathbf{\Phi} + \mathbf{A})^{-1}\mathbf{\Phi}^\top \mathbf{t},$$

where the noise variance σ^2 is assumed to be known.

In the particular case where $\alpha_m = \alpha_0$ for all m (and σ is fixed), the MAP solution is equivalent to **ridge regression**, since we have $\alpha = \alpha_0\sigma^2$.

In order to infer the amount of noise, one has to use the Expectation-Maximisation algorithm as $\mathbf{w}$ and σ^2 are coupled (see later).

The log-posterior is given by

$$\ln p(\mathbf{w}|\mathbf{t}) = -\frac{N}{2}\ln 2\pi - \frac{N}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}(\mathbf{t}-\mathbf{y})^\top(\mathbf{t}-\mathbf{y}) - \frac{M+1}{2}\ln 2\pi + \frac{1}{2}\ln|\mathbf{A}| - \frac{1}{2}\mathbf{w}^\top\mathbf{A}\mathbf{w} - \ln Z.$$

Hence, this leads to

$$\frac{d\ln p(\mathbf{w}|\mathbf{t})}{d\mathbf{w}} = 0 \quad \Rightarrow \quad \frac{1}{\sigma^2}\boldsymbol{\Phi}^\top(\mathbf{t}-\boldsymbol{\Phi}\mathbf{w}) - \mathbf{A}\mathbf{w} = 0$$

$$\Leftrightarrow \quad \frac{1}{\sigma^2}\boldsymbol{\Phi}^\top\mathbf{t} = \frac{1}{\sigma^2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\mathbf{w} + \mathbf{A}\mathbf{w}.$$

Is the MAP solution a good solution?

- Overfitting and the model selection problem (i.e. the choice of the number of prototypes) are solved by limiting the **effective model complexity**.
- It might be difficult to deal with (very) large data sets.
- The better (smoother) solution is at the cost of **additional hyperparameters**, which can only be set by resampling.
- MAP makes predictions based on **point estimates** as ML; the uncertainty on the parameters is not taken into account when making predictions:

$$p(t|\mathbf{t}) \approx p(t|\mathbf{w}_{\mathrm{MAP}}) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}_{\mathrm{MAP}}), \sigma^2).$$

- The MAP solution depends on the **parametrisation** of the prior.

Type II ML (or evidence maximisation)

We view some parameters as **latent**, i.e. unobserved, random variables.

In order to deal properly with the uncertainty, we want to **integrate them out** before estimating the remaining parameters by ML or making predictions.

Assume the random variable $\mathbf{x}$ is observed and the variable $\mathbf{z}$ is unobserved. Let us denote the (remaining) parameters by $\boldsymbol{\theta}$.

1. The full **predictive distribution** is approximated by

$$p(\mathbf{x}_*|\mathbf{x}) \approx \int p(\mathbf{x}_*|\mathbf{z},\boldsymbol{\theta})\; p(\mathbf{z}|\mathbf{x},\boldsymbol{\theta})\; d\mathbf{z}.$$

   where $p(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}) = \frac{p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})}{p(\mathbf{x}|\boldsymbol{\theta})}$ is the posterior.

2. The parameters are estimated by maximising (e.g. by gradient descent) the **marginal likelihood** (or evidence)[1]:

$$\boldsymbol{\theta}_{\mathrm{ML2}} = \operatorname*{argmax}_{\boldsymbol{\theta}}\; p(\mathbf{x}|\boldsymbol{\theta}),$$

   where $p(\mathbf{x}|\boldsymbol{\theta}) = \int p(\mathbf{x},\mathbf{z}|\boldsymbol{\theta})\; d\mathbf{z}$.

[1] Type II ML might be more sensitive to model mis-specification than resampling techniques based on the *pseudo-likelihood* (see Section 4.8 in *Wahba*, 1990).

The Expectation-Maximisation (EM) algorithm in a nutshell

The EM algorithm maximises a **lower bound** of the log-marginal likelihood in presence of latent variables.

Using *Jensen's inequality*, we get for a distribution $q(\mathbf{z})$ within a tractable family:

$$\begin{aligned} \ln p(\mathbf{x}|\boldsymbol{\theta}) &= \ln \int p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) d\mathbf{z} \\ &\geqslant \int q(\mathbf{z}) \ln \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})} d\mathbf{z} \\ &\equiv -\mathcal{F}(q, \boldsymbol{\theta}). \end{aligned}$$

The **variational free energy** $\mathcal{F}(q, \boldsymbol{\theta})$ can be decomposed in two different ways:

$$\begin{aligned} -\mathcal{F}(q, \boldsymbol{\theta}) &= \ln p(\mathbf{x}|\boldsymbol{\theta}) - \mathrm{KL}[q(\mathbf{z}) \| p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})], && \text{(E step)} \\ -\mathcal{F}(q, \boldsymbol{\theta}) &= \langle \ln p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) \rangle_{q(\mathbf{z})} + \mathrm{H}[q(\mathbf{z})]. && \text{(M step)} \end{aligned}$$

The EM algorithm maximises the lower bound by alternatively minimising the KL for fixed parameters (E step) and maximising the expected complete log-likelihood for a fixed $q(\mathbf{z})$ (M step).

By construction, the EM algorithm ensures a **monotonic** increase of the bound.

EM for linear regressors

We view $\mathbf{w}$ as a latent (unobserved) variable on which an isotropic **Gaussian prior** is imposed:

$$\mathbf{w}|\alpha \sim \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I}_{M+1}).$$

The goal is to learn the noise variance σ^2 and the scale parameter α.

The E step is exact as the posterior on $\mathbf{w}$ is tractable (conjugate prior):

$$\mathbf{w}|\mathbf{t}, \sigma, \alpha \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{w}, \boldsymbol{\Sigma}_\mathbf{w}),$$

where $\boldsymbol{\mu}_\mathbf{w} \equiv \sigma^{-2}\boldsymbol{\Sigma}_\mathbf{w}\boldsymbol{\Phi}^\top\mathbf{t}$ and $\boldsymbol{\Sigma}_\mathbf{w} \equiv (\sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \alpha\mathbf{I}_{M+1})^{-1}$.

For a Gaussian distribution the mode is equal to the mean. Hence, the estimate for $\mathbf{w}$ is equal to the one obtained before, i.e. $\langle\mathbf{w}\rangle = \boldsymbol{\mu}_\mathbf{w} = \mathbf{w}_{\mathrm{MAP}}$.

The M step is given by

$$\sigma^2_{\mathrm{ML2}} \leftarrow \operatorname*{argmax}_{\sigma^2} \langle \ln p(\mathbf{t}, \mathbf{w}|\sigma, \alpha)\rangle = \frac{\|\mathbf{t} - \boldsymbol{\Phi}\boldsymbol{\mu}_\mathbf{w}\|^2 + \mathrm{tr}\{\boldsymbol{\Phi}\boldsymbol{\Sigma}_\mathbf{w}\boldsymbol{\Phi}^\top\}}{N},$$

$$\alpha_{\mathrm{ML2}} \leftarrow \operatorname*{argmax}_{\alpha} \langle \ln p(\mathbf{t}, \mathbf{w}|\sigma, \alpha)\rangle = \frac{M+1}{\boldsymbol{\mu}_\mathbf{w}^\top\boldsymbol{\mu}_\mathbf{w} + \mathrm{tr}\{\boldsymbol{\Sigma}_\mathbf{w}\}}.$$

The posterior is given by (completing the square)

$$
\begin{aligned}
p(\mathbf{w}|\mathbf{t},\sigma,\alpha) &\propto e^{-\frac{1}{2\sigma^2}(\mathbf{t}-\boldsymbol{\Phi}\mathbf{w})^\top(\mathbf{t}-\boldsymbol{\Phi}\mathbf{w})}\; e^{-\frac{\alpha}{2}\mathbf{w}^\top\mathbf{w}} \\
&\propto e^{-\frac{1}{2}(\mathbf{w}^\top(\sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}+\alpha\mathbf{I}_{M+1})\mathbf{w}-2\sigma^{-2}\mathbf{t}^\top\boldsymbol{\Phi}\mathbf{w})} \\
&\propto e^{-\frac{1}{2}(\mathbf{w}^\top\boldsymbol{\Sigma}_\mathbf{w}^{-1}\mathbf{w}-2\boldsymbol{\mu}_\mathbf{w}^\top\boldsymbol{\Sigma}_\mathbf{w}^{-1}\mathbf{w})} \\
&\propto e^{-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu}_\mathbf{w})^\top\boldsymbol{\Sigma}_\mathbf{w}^{-1}(\mathbf{w}-\boldsymbol{\mu}_\mathbf{w})}.
\end{aligned}
$$

The expected complete log-likelihood is given by

$$
\begin{aligned}
\langle \ln p(\mathbf{t},\mathbf{w}|\sigma,\alpha)\rangle &= -\frac{N}{2}\ln 2\pi - \frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\left\langle(\mathbf{t}-\mathbf{y})^\top(\mathbf{t}-\mathbf{y})\right\rangle \\
&\quad - \frac{M+1}{2}\ln 2\pi + \frac{M+1}{2}\ln\alpha - \frac{\alpha}{2}\langle\mathbf{w}^\top\mathbf{w}\rangle \\
&= -\frac{N}{2}\ln 2\pi - \frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{t}-\boldsymbol{\Phi}\langle\mathbf{w}\rangle)^\top(\mathbf{t}-\boldsymbol{\Phi}\langle\mathbf{w}\rangle) \\
&\quad - \frac{1}{2\sigma^2}\mathrm{tr}\{\boldsymbol{\Phi}\boldsymbol{\Sigma}_\mathbf{w}\boldsymbol{\Phi}^\top\} - \frac{M+1}{2}\ln 2\pi + \frac{M+1}{2}\ln\alpha - \frac{\alpha}{2}\langle\mathbf{w}^\top\mathbf{w}\rangle.
\end{aligned}
$$

Taking the derivative wrt σ^2 and α, and equating to zero leads to the desired updates.

Predictive distributions

We are not only interested in the optimal predictions, but also in the best approximation of the full predictive distribution.

The predictive distributions for the ML and the type II ML solutions are given by

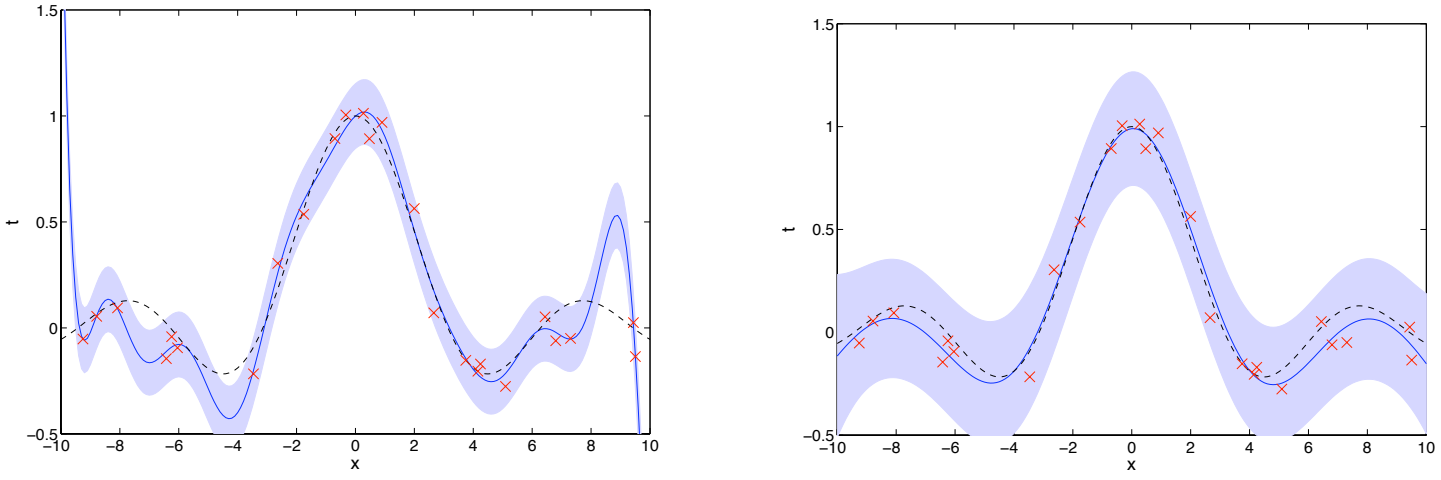
$$p(t|\mathbf{t}) \approx p(t|\mathbf{w}_{\mathrm{ML}}, \sigma_{\mathrm{ML}}) = \mathcal{N}(\mathbf{w}_{\mathrm{ML}}^{\top}\boldsymbol{\phi}(\mathbf{x}), \sigma_{\mathrm{ML}}^2),$$
$$p(t|\mathbf{t}) \approx p(t|\mathbf{t}, \sigma_{\mathrm{ML2}}, \alpha_{\mathrm{ML2}}) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{w}}^{\top}\boldsymbol{\phi}(\mathbf{x}), \sigma_{\mathrm{ML2}}^2 + \boldsymbol{\phi}^{\top}(\mathbf{x})\boldsymbol{\Sigma}_{\mathbf{w}}\boldsymbol{\phi}(\mathbf{x})).$$

In the case of type II ML, the **predictive variance** has two components:

- The noise on the data.
- The uncertainty associated to the parameters.

Example revisited

We compare the solutions on the sinc example with $N = 25$, $\sigma = 0.1$ and $\lambda_m = 1/9$ for all m. We show the mean and the error bars (± 3 std):

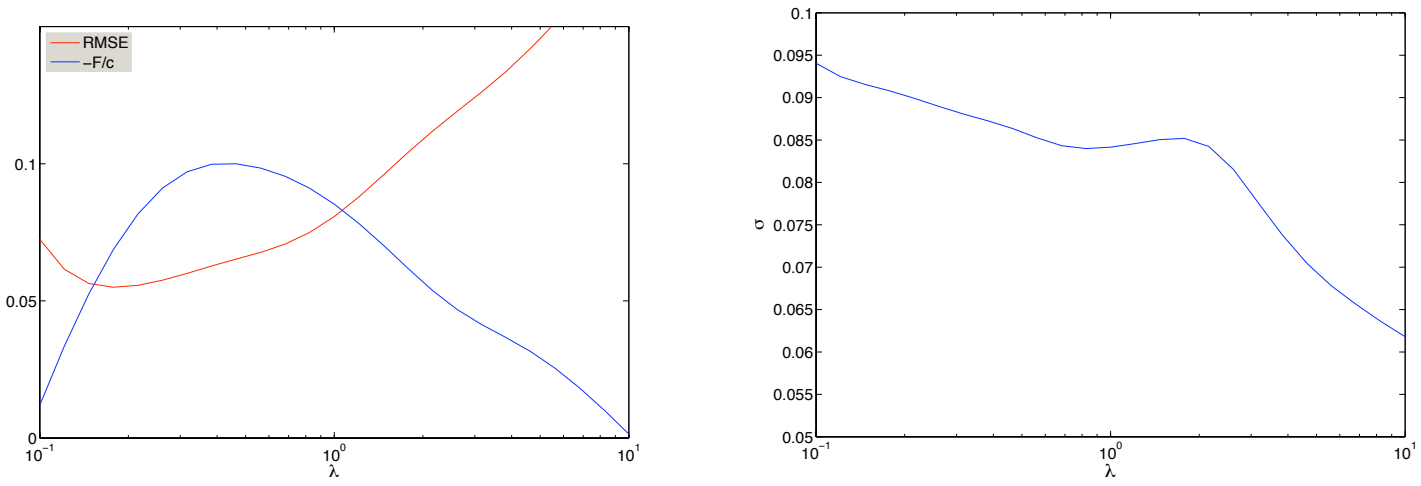


(a) ML. (b) ML2.

Figure: (a) ML solution: $\sigma_{\mathrm{ML}} = 0.05$. (b) Type II ML solution: $\sigma_{\mathrm{ML2}} = 0.08$ and $\alpha_{\mathrm{ML2}} = 1.15$. (Target function: dashed; observations: crosses.)

Is the type II ML solution a good solution?

- The solution **avoids overfitting** by taking some uncertainty into account.
- Predictions provide **error bars** by integrating out **w**.
- The lower bound is moderately suitable for selecting the **kernel width**:



(a) (b)

Figure: (a) Root mean square error (RMSE) and normalised lower bound ($-\mathcal{F}/c$) versus the kernel width λ. (b) Noise standard deviation versus λ.

- Will we gain something if we take the uncertainty of the hyperparameters into acount and be agnostic at a **higher level**?

Variational Bayes (VEM) in a nutshell

Variational Bayes **generalises EM** by viewing all parameters as random variables.

Using *Jensen's inequality*, we obtain a lower bound on the log-marginal likelihood:

$$\begin{aligned}
\ln p(\mathbf{x}) &= \ln \iint p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta}) \, d\mathbf{z} \, d\boldsymbol{\theta} \\
&\geqslant \iint q(\mathbf{z}) q(\boldsymbol{\theta}) \ln \frac{p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta})}{q(\mathbf{z}) q(\boldsymbol{\theta})} \, d\mathbf{z} \, d\boldsymbol{\theta} \\
&= \ln p(\mathbf{x}) - \mathrm{KL}[q(\mathbf{z}) q(\boldsymbol{\theta}) \| p(\mathbf{z}, \boldsymbol{\theta} | \mathbf{x})] \\
&\equiv -\mathcal{F}(q(\mathbf{z}), q(\boldsymbol{\theta})).
\end{aligned}$$

For tractability it is assumed that the variational posterior **factorises** (at least) between the latent variables and the parameters given the observations.

When a fully factorised posterior is assumed, one talks about **mean field**.

Assuming a factorised form corresponds to **neglecting the correlations** between dependent variables and thus prevents transmitting uncertainty.

Considering $\mathrm{KL}[q\|p]$ leads a more **compact** posterior as it is zero forcing.

Variational Bayes in a nutshell (continued)

The variational bound can be maximised by alternating between the following updates:

$$q(\mathbf{z}) \propto e^{\langle \ln p(\mathbf{x},\mathbf{z}|\boldsymbol{\theta}|)\rangle_{q(\boldsymbol{\theta})}}, \qquad \text{(VE step)}$$

$$q(\boldsymbol{\theta}) \propto e^{\langle \ln p(\mathbf{x},\mathbf{z}|\boldsymbol{\theta})\rangle_{q(\mathbf{z})}} p(\boldsymbol{\theta}), \qquad \text{(VM step)}$$

In practice, the priors are chosen to be **conjugate** to the likelihood such that updating the posterior simply consists in updating the hyperparameters.

By construction, VEM ensures a **monotonic** increase of the bound.

The **variational lower bound** can be evaluated using

$$-\mathcal{F}(q(\mathbf{z}), q(\boldsymbol{\theta})) = \langle \ln p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta})\rangle_{q(\mathbf{z})q(\boldsymbol{\theta})} + \mathrm{H}[q(\mathbf{z})] + \mathrm{H}[q(\boldsymbol{\theta})].$$

The **predictive distribution** is approximated as follows:

$$p(\mathbf{x}_*|\mathbf{x}) \approx \int p(\mathbf{x}_*|\mathbf{z}, \boldsymbol{\theta}) q(\mathbf{z}) q(\boldsymbol{\theta}) d\mathbf{z} d\boldsymbol{\theta}.$$

First, we show that VE maximises the lower bound (i.e. minimises the free energy) for fixed $q(\boldsymbol{\theta})$:

$$\begin{aligned}
\mathcal{F} &= -\int q(\mathbf{z}) \left\langle \ln p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) \right\rangle_{q(\boldsymbol{\theta})} d\mathbf{z} - \mathrm{H}[q(\mathbf{z})] + \text{const.} \\
&= -\int q(\mathbf{z}) \ln \frac{e^{\left\langle \ln p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) \right\rangle_{q(\boldsymbol{\theta})}}}{q(\mathbf{z})} d\mathbf{z} + \text{const.} \\
&= \mathrm{KL}\left[q(\mathbf{z}) \middle\| \tfrac{1}{Z} e^{\left\langle \ln p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) \right\rangle_{q(\boldsymbol{\theta})}} \right] + \text{const.}
\end{aligned}$$

Second, we show that VM maximises the lower bound for fixed $q(\mathbf{z})$:

$$\begin{aligned}
\mathcal{F} &= -\int q(\boldsymbol{\theta}) \left\langle \ln p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta}) \right\rangle_{q(\mathbf{z})} d\boldsymbol{\theta} - \mathrm{H}[q(\boldsymbol{\theta})] + \text{const.} \\
&= -\int q(\boldsymbol{\theta}) \ln \frac{e^{\left\langle \ln p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta}) \right\rangle_{q(\mathbf{z})}}}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} + \text{const.} \\
&= \mathrm{KL}\left[q(\boldsymbol{\theta}) \middle\| \tfrac{1}{Z} e^{\left\langle \ln p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) \right\rangle_{q(\mathbf{z})}} p(\boldsymbol{\theta}) \right] + \text{const.}
\end{aligned}$$

Bayesian linear models for regression

For simplicity, the prior on the weights is assumed to be isotropic Gaussian:

$$\mathbf{w}|\alpha \sim \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I}_{M+1}),$$

To force non-negative value, we impose **Gamma priors** on α and $\tau = \sigma^{-2}$:

$$\alpha \sim \mathcal{G}(a_0, b_0),$$
$$\tau \sim \mathcal{G}(c_0, d_0).$$

The Gamma is conjugate to the Gaussian.

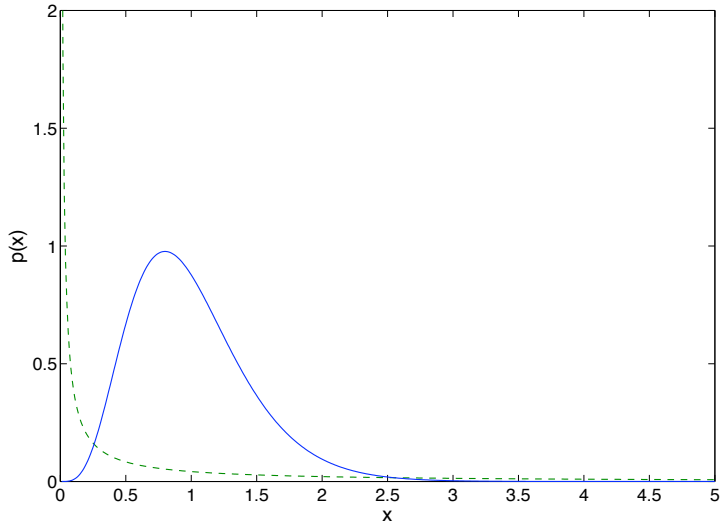


Figure: Gamma distribution for various values of the scale and the shape parameter.

Variational inference for Bayesian linear regressors

The complete log-likelihood is given by

$$p(\mathbf{t}, \mathbf{w}, \tau, \alpha) = p(\mathbf{t}|\mathbf{w}, \tau)p(\mathbf{w}|\alpha)p(\alpha)p(\tau).$$

It is assumed that the variational posterior fully factorises, such that

$$q(\mathbf{w}, \alpha, \tau) = q(\mathbf{w})q(\alpha)q(\tau).$$

Applying the variational framework leads to

$$\begin{aligned} q(\mathbf{w}) &= \mathcal{N}(\bar{\boldsymbol{\mu}}_{\mathbf{w}}, \bar{\boldsymbol{\Sigma}}_{\mathbf{w}}), \\ q(\alpha) &= \mathcal{G}(a, b), \\ q(\tau) &= \mathcal{G}(c, b), \end{aligned}$$

where the special quantities are defined as

$$\begin{aligned} &\bar{\boldsymbol{\mu}}_{\mathbf{w}} \equiv \langle\tau\rangle \bar{\boldsymbol{\Sigma}}_{\mathbf{w}} \boldsymbol{\Phi}^{\top} \mathbf{t}, \quad \bar{\boldsymbol{\Sigma}}_{\mathbf{w}} \equiv (\langle\tau\rangle \boldsymbol{\Phi}^{\top}\boldsymbol{\Phi} + \langle\alpha\rangle \mathbf{I}_{M+1})^{-1}, \\ &a \equiv \frac{M+1}{2} + a_0, \quad b \equiv \frac{\bar{\boldsymbol{\mu}}_{\mathbf{w}}^{\top}\bar{\boldsymbol{\mu}}_{\mathbf{w}} + \mathrm{tr}\{\bar{\boldsymbol{\Sigma}}_{\mathbf{w}}\}}{2} + b_0, \\ &c \equiv \frac{N}{2} + c_0, \quad d \equiv \frac{\|\mathbf{t} - \boldsymbol{\Phi}\bar{\boldsymbol{\mu}}_{\mathbf{w}}\|^2 + \mathrm{tr}\{\boldsymbol{\Phi}\bar{\boldsymbol{\Sigma}}_{\mathbf{w}}\boldsymbol{\Phi}^{\top}\}}{2} + d_0, \end{aligned}$$

where $\langle\tau\rangle = c/d$ and $\langle\alpha\rangle = a/b$.

The variational posterior for $\mathbf{w}$ is given by

$$\begin{aligned} q(\mathbf{w}) &\propto e^{\langle \ln p(\mathbf{t},\mathbf{w}|\alpha,\tau)\rangle_{q(\alpha)q(\tau)}} \\ &= e^{-\frac{1}{2}(\mathbf{w}-\bar{\boldsymbol{\mu}}_{\mathbf{w}})^{\top}\bar{\boldsymbol{\Sigma}}_{\mathbf{w}}^{-1}(\mathbf{w}-\bar{\boldsymbol{\mu}}_{\mathbf{w}})}. \end{aligned}$$

The variational posterior for α is given by

$$\begin{aligned} q(\alpha) &\propto e^{\langle \ln p(\mathbf{t},\mathbf{w}|\alpha,\tau)\rangle_{q(\mathbf{w})q(\tau)}}\ p(\alpha) \\ &\propto e^{(a-1)\ln\alpha - b\alpha}, \end{aligned}$$

where $a \equiv \frac{M+1}{2} + a_0$ and $b \equiv \frac{\langle \mathbf{w}^{\top}\mathbf{w}\rangle}{2} + b_0$.

The variational posterior for α is given by

$$\begin{aligned} q(\tau) &\propto e^{\langle \ln p(\mathbf{t},\mathbf{w}|\alpha,\tau)\rangle_{q(\mathbf{w})q(\alpha)}}\ p(\tau) \\ &\propto e^{(c-1)\ln\tau - d\tau}. \end{aligned}$$

where $c \equiv \frac{N}{2} + c_0$ and $d \equiv \frac{\langle (\mathbf{t}-\boldsymbol{\Phi}\mathbf{w})^{\top}(\mathbf{t}-\boldsymbol{\Phi}\mathbf{w})\rangle}{2} + d_0$.

Variational *vs* type II ML

In practice, the variational approach consists in updating the parameters of the posteriors in turn.

The updates for the posterior mean and the posterior covariance of $\mathbf{w}$ have a **similar form** as before.

The estimates of σ^2 and α are replaced by their expectation:

$$\langle \tau \rangle^{-1} = \frac{\|\mathbf{t} - \mathbf{\Phi}\bar{\boldsymbol{\mu}}_{\mathbf{w}}\|^2 + \mathrm{tr}\{\mathbf{\Phi}\bar{\mathbf{\Sigma}}_{\mathbf{w}}\mathbf{\Phi}^\top\} + 2d_0}{N + 2c_0},$$

$$\langle \alpha \rangle = \frac{M + 1 + 2a_0}{\bar{\boldsymbol{\mu}}_{\mathbf{w}}^\top \bar{\boldsymbol{\mu}}_{\mathbf{w}} + \mathrm{tr}\{\bar{\mathbf{\Sigma}}_{\mathbf{w}}\} + 2b_0}.$$

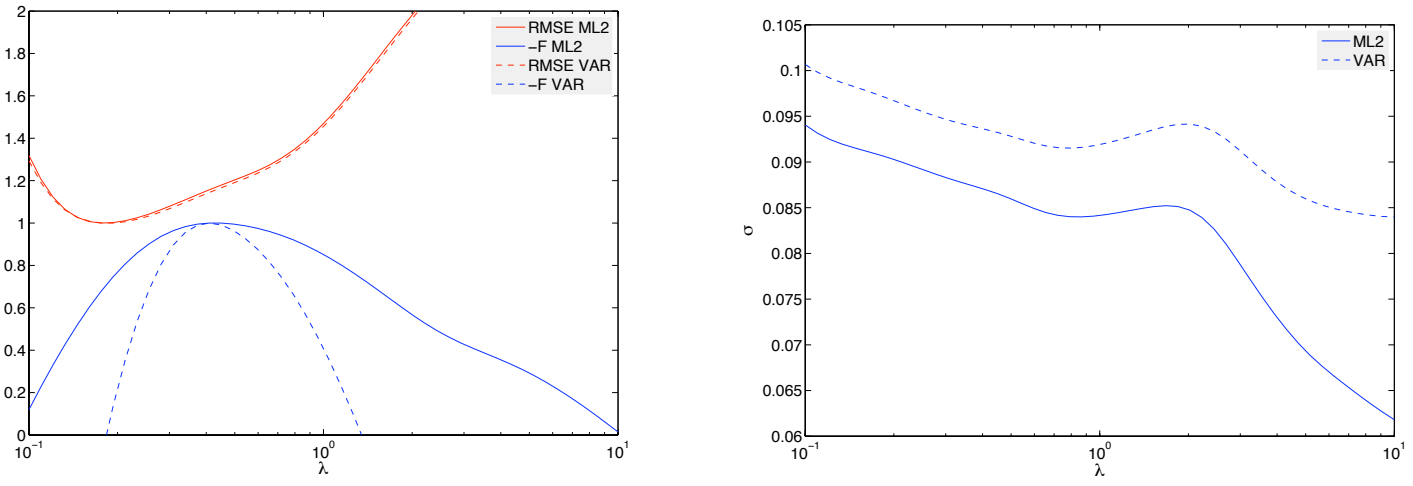
For flat priors, these estimates will give very similar results as type II ML.

The **predictive distribution** is also very similar in form:

$$p(t|\mathbf{t}) \approx \int p(t|\mathbf{w}, \langle \tau \rangle^{-1})\ q(\mathbf{w})\ d\mathbf{w} = \mathcal{N}(\bar{\boldsymbol{\mu}}_{\mathbf{w}}^\top \boldsymbol{\phi}(\mathbf{x}), \langle \tau \rangle^{-1} + \boldsymbol{\phi}^\top(\mathbf{x})\bar{\mathbf{\Sigma}}_{\mathbf{w}}\boldsymbol{\phi}(\mathbf{x})).$$

Variational *vs* type II ML (continued)

The variational approach does not provide a better selection criterion for the kernel width, but leads to a better estimate of the observation noise!



(a) (b)

Figure: Comparison of the quality of the type II ML solution and the variational solution. (a) Normalised root mean square error (RMSE) and normalised lower bound ($-\mathcal{F}$) versus the kernel width λ. (b) Noise standard deviation σ versus λ. Note that the true value of σ is 0.1.

Selecting the number of the basis functions

Why seeking sparse solutions?

Placing a basis function on each training data ($M = N$) rapidly becomes infeasible in practice:

- Training becomes slow (cf. matrix $\mathbf{\Sigma_w} \in \mathbb{R}^{(M+1)\times(M+1)}$ to invert).
- Even if training off-line making predictions may be too expensive.
- Excessive memory usage as the design matrix grows quadratically with M.

Sparse solutions lead to better generalisation and are **fast** (e.g. SVM).

Simple heuristics based on resampling:

- Subset selection (randomly or based on an information theoretic criterion).
- Vector quantisation or clustering.
- ...

These approaches are unsupervised and thus **suboptimal**.

Sparse Bayesian linear models for regression

A systematic approach is to build a **hierarchical model** such that the effective prior on **w** is **sparsity inducing**, i.e. most weights are driven to zero:

- *Sparse Bayesian learning and the relevance vector machines* (Tipping, 2001)
- *Adaptive sparseness for supervised learning* (Figueiredo, 2003)
- *Comparing the effects of different weight distributions on finding sparse representations* (Wipf and Rao, 2006)
- *Bayesian adaptive lassos with non-convex penalization* (Griffin and Brown, 2007)
- ...

These methods lead to very sparse solutions (more sparse than standard SVMs).

Relevance vector machines

Consider a Gaussian prior with a different scale parameter α_m for each w_m, along with a different hyperprior:

$$\mathbf{w}|\alpha_0, \ldots, \alpha_M \sim \mathcal{N}(0, \mathbf{A}^{-1}) = \prod_{m=0}^{M} \mathcal{N}(0, \alpha_m^{-1}),$$

$$(\alpha_0, \ldots, \alpha_M) \sim \prod_{m=0}^{M} \mathcal{G}(a_m, b_m).$$

Integrating out α_m leads to an **effective prior** on w_m which corresponds to a Student-t:

$$\begin{aligned} p(w_m) &= \int \mathcal{N}(0, \alpha_m^{-1})\ \mathcal{G}(a_m, b_m)\ d\alpha_m \\ &= \mathcal{S}(0, a_m/b_m, 2a_m), \end{aligned}$$

for all m. The Student-t distribution approximates the Laplace distribution (or double exponential), which corresponds to L_1-regularisation (LASSO).

The Student-t prior is **peaked around zero** and has **fat tails**.

Sparsity inducing priors

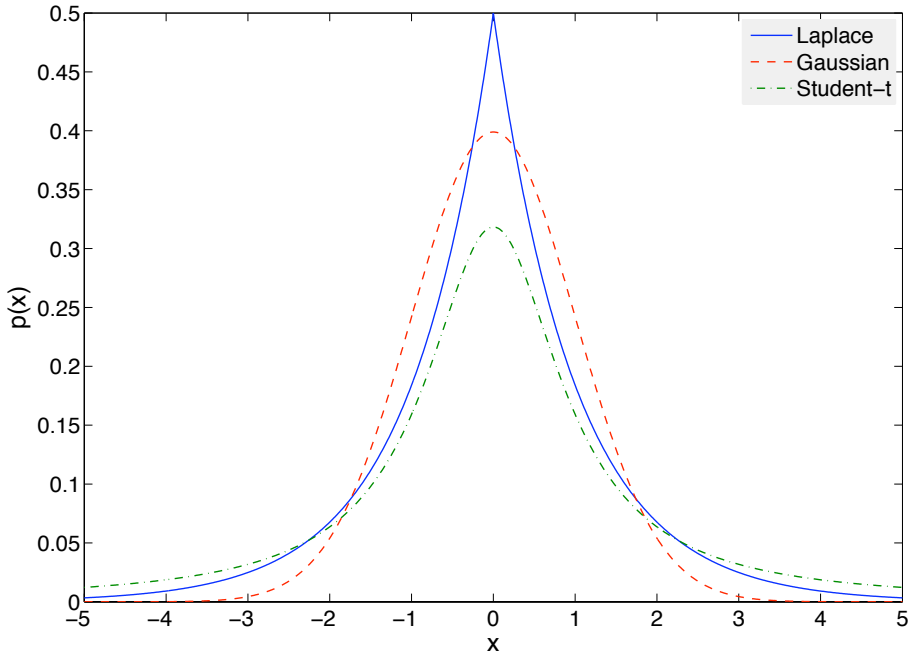


Figure: Two sparsity inducing priors compared to the Gaussian.

References

- Christopher M. Bishop: Pattern Recognition and Machine Learning. Springer, 2006.
- Trevor Hastie, Robert Tibshirani and Jerome Friedman: The Elements of Statistical Learning. Springer, 2001.
- John Shawe-Taylor and Nello Cristianini: Kernel Methods for Pattern Analysis. Cambridge University Press, 2004.