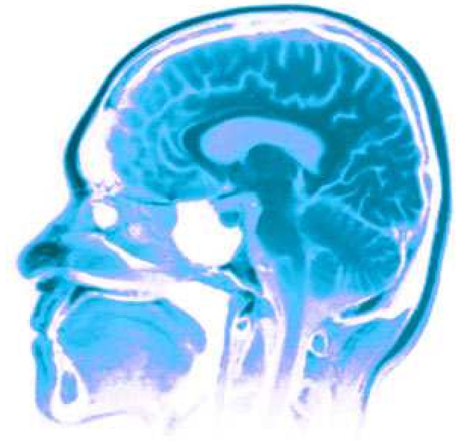
CPSC340



Ridge regression and regularization

Nando de Freitas
October, 2012
University of British Columbia

Outline of the lecture

This lecture introduces regularization and Bayesian learning for the linear Gaussian model. The goal is for you to:

- ❑ Learn how to derive **ridge regression**.
- ❑ Understand the trade-off of fitting the data and **regularizing** it.
- ❑ Derive the **Bayesian** estimates for linear regression.

Regularization

All the answers so far are of the form

$$\widehat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

They require the inversion of $\mathbf{X}^T\mathbf{X}$. This can lead to problems if the system of equations is poorly conditioned. A solution is to add a small element to the diagonal:

$$\widehat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X} + \delta^2 I_d)^{-1}\mathbf{X}^T\mathbf{y}$$

This is the ridge regression estimate. It is the solution to the following **regularised quadratic cost function**

$$J(\boldsymbol{\theta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \delta^2\boldsymbol{\theta}^T\boldsymbol{\theta}$$

Derivation

$$\frac{\partial}{\partial\theta} J(\theta) = \frac{\partial}{\partial\theta}\left\{ (y - X\theta)^T (y - X\theta) + \delta^2 \theta I \theta \right\}$$

identity matrix

$$= \frac{\partial}{\partial\theta}\left\{ y^T y - 2y^T X\theta + \theta^T X^T X\theta + \theta^T (\delta^2 I)\theta \right\}$$

$$= -2X^T y + 2X^T X\theta + 2\delta^2 I \theta$$

$$= -2X^T y + 2(X^T X + \delta^2 I)\theta$$

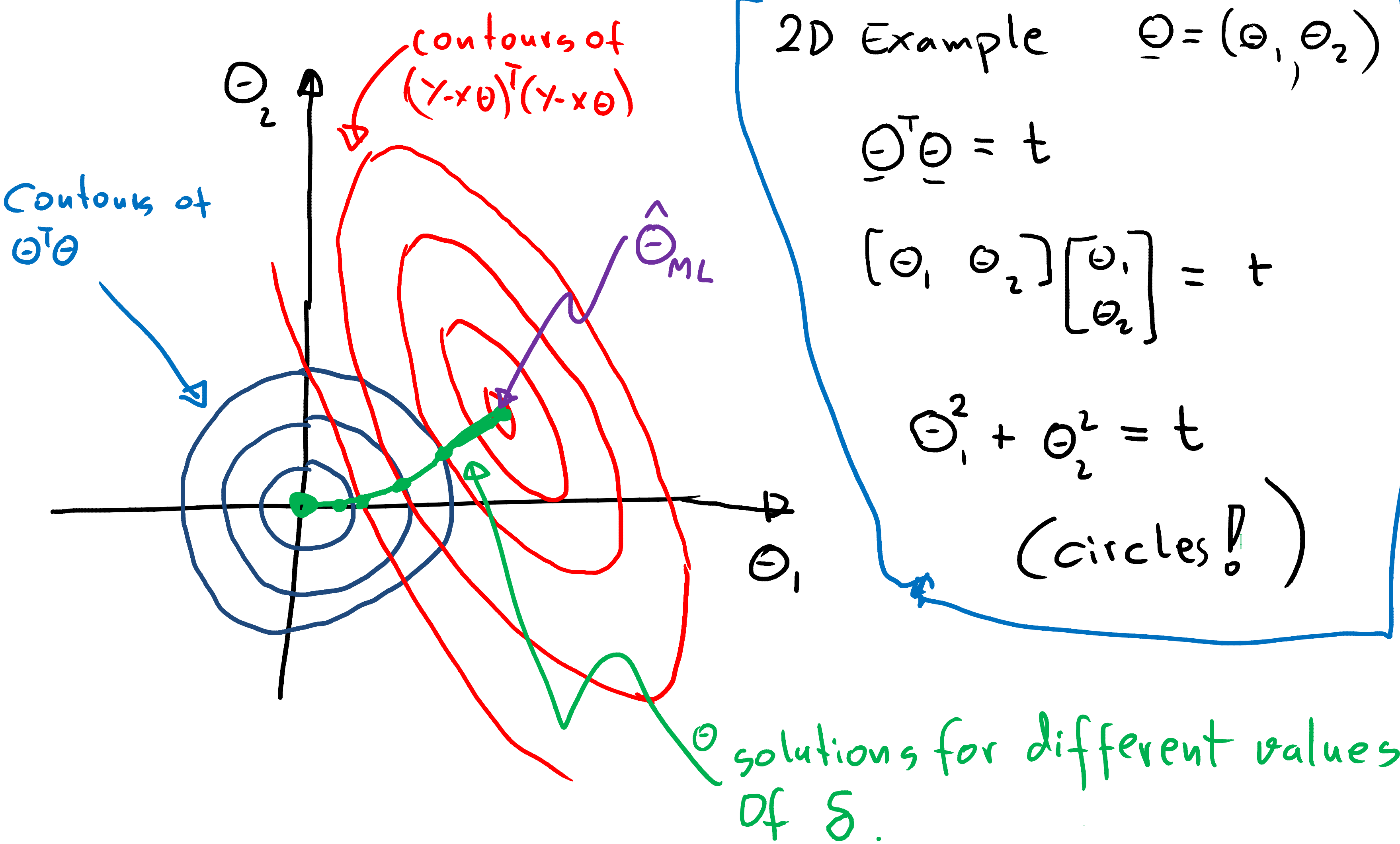
Equating to zero, yields

$$\hat{\theta}_{ridge} = (X^T X + \delta^2 I)^{-1} X^T y$$

Ridge regression as constrained optimization

$$J(\boldsymbol{\theta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \delta^2\boldsymbol{\theta}^T\boldsymbol{\theta}$$

$$\min_{\boldsymbol{\theta}\,:\;\boldsymbol{\theta}^T\boldsymbol{\theta} \leq t(\delta)} \{(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})\}$$



2D Example $\underline{\theta} = (\theta_1, \theta_2)$

$\underline{\theta}^T\underline{\theta} = t$

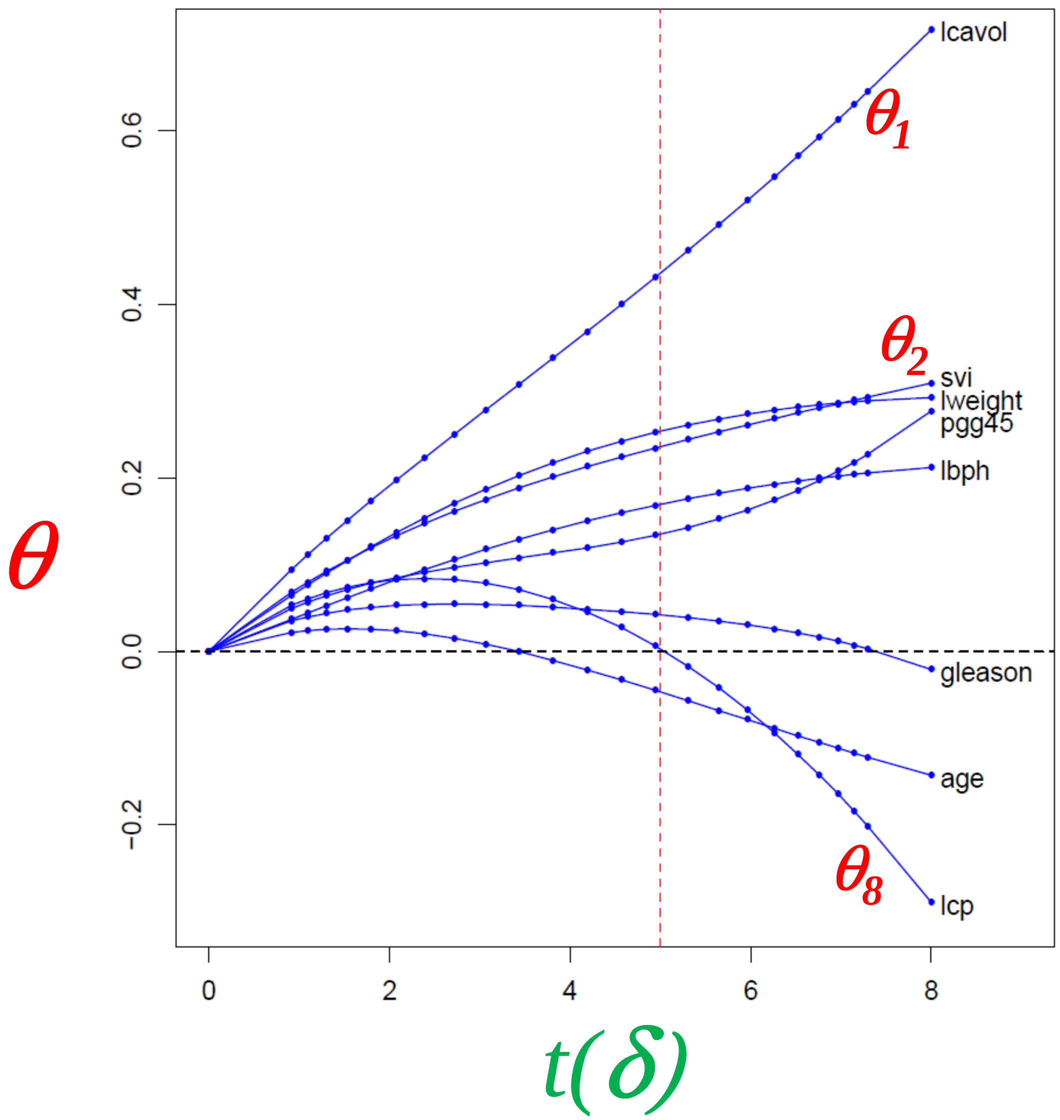
$\begin{bmatrix}\theta_1 & \theta_2\end{bmatrix}\begin{bmatrix}\theta_1 \\ \theta_2\end{bmatrix} = t$

$\theta_1^2 + \theta_2^2 = t$

(circles!)

Regularization paths

As δ increases, $t(\delta)$ decreases and each θ_i goes to zero.



[Hastie, Tibshirani & Friedman book]

Ridge, feature selection, shrinkage and weight decay

Large values of $\boldsymbol{\theta}$ are penalised. We are *shrinking* $\boldsymbol{\theta}$ towards zero. This can be used to carry out *feature weighting*. An input $x_{i,d}$ weighted by a small θ_d will have less influence on the ouptut y_i. This penalization with a regularizer is also known as weight decay in the neural networks literature.

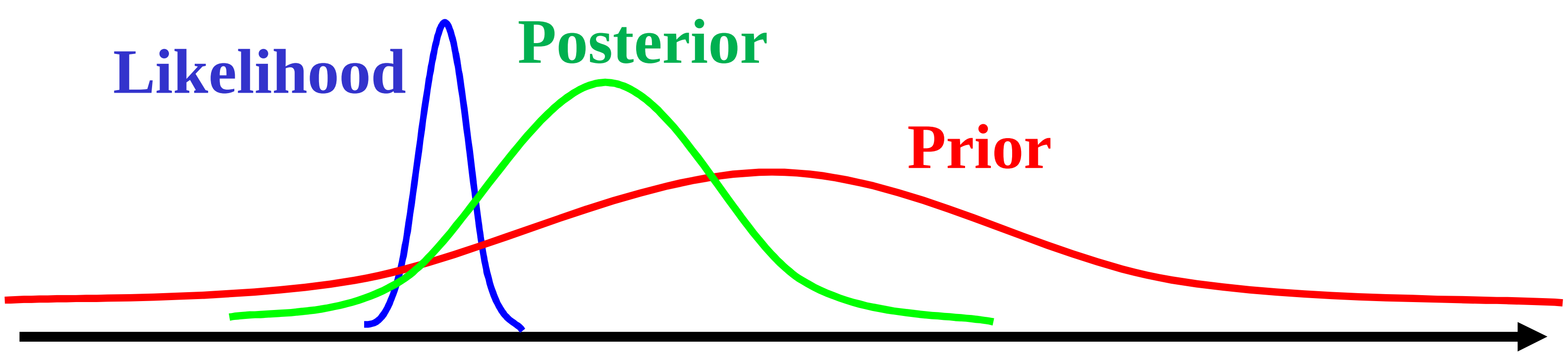
Note that shrinking the bias term $\boldsymbol{\theta}_1$ is undesirable. To keep the notation simple, we will assume that the mean of $\mathbf{y}$ has been subtracted from $\mathbf{y}$. This mean is indeed our estimate $\widehat{\boldsymbol{\theta}_1}$.

Bayesian linear regression

The likelihood is a Gaussian, $\mathcal{N}(\mathbf{y}|\mathbf{X}\boldsymbol{\theta}, \sigma^2\mathbf{I}_n)$. The conjugate prior is also a Gaussian, which we will denote by $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\theta}_0, \mathbf{V}_0)$.

Using Bayes rule for Gaussians, the posterior is given by

$$\begin{aligned}
p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}, \sigma^2) &\propto \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\theta}_0, \mathbf{V}_0)\mathcal{N}(\mathbf{y}|\mathbf{X}\boldsymbol{\theta}, \sigma^2\mathbf{I}_n) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\theta}_n, \mathbf{V}_n) \\
\boldsymbol{\theta}_n &= \mathbf{V}_n\mathbf{V}_0^{-1}\boldsymbol{\theta}_0 + \frac{1}{\sigma^2}\mathbf{V}_n\mathbf{X}^T\mathbf{y} \\
\mathbf{V}_n^{-1} &= \mathbf{V}_0^{-1} + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}
\end{aligned}$$



Bayesian linear regression

Assume σ^2 is known.

$$P(\theta \mid X, Y, \sigma^2) \propto P(Y \mid X, \theta, \sigma^2)\, P(\theta)$$

$$\propto e^{-\frac{1}{2}(Y - X\theta)^T (\sigma^2 I)^{-1} (Y - X\theta)}\; e^{-\frac{1}{2}(\theta - \theta_0)^T V_0^{-1} (\theta - \theta_0)}$$

$$= e^{-\frac{1}{2}\left\{ Y^T(\sigma^2 I)^{-1} Y - 2Y^T(\sigma^2 I)^{-1} X\theta + \theta^T X^T (\sigma^2 I)^{-1} X\theta + \theta^T V_0^{-1}\theta + \theta_0^T V_0^{-1}\theta_0 - 2\theta_0^T V_0^{-1}\theta \right\}}$$

Call this V_n^{-1}:

$$= e^{-\frac{1}{2}\left\{ \text{const} + \theta^T \underbrace{\left(X^T(\sigma^2 I)^{-1}X + V_0^{-1}\right)}_{V_n^{-1}} \theta - 2\left(Y^T(\sigma^2 I)^{-1} X + \theta_0^T V_0^{-1}\right)\theta \right\}}$$

$$= e^{-\frac{1}{2}\left\{ \text{const} + \theta^T V_n^{-1}\theta - 2\left(\frac{Y^T X}{\sigma^2} + \theta_0^T V_0^{-1}\right)\theta \right\}}$$

$$= e^{-\frac{1}{2}\left\{ \text{const} + \theta^T V_n^{-1}\theta - 2\theta_n^T V_n^{-1}\theta + 2\theta_n^T V_n^{-1}\theta - 2\left(\frac{Y^T X}{\sigma^2} + \theta_0^T V_0^{-1}\right)\theta \right\}}$$

$$= e^{-\frac{1}{2}\left\{ \text{const}_2 + (\theta - \theta_n)^T V_n^{-1}(\theta - \theta_n) + 2\left[\theta_n^T V_n^{-1} - \frac{Y^T X}{\sigma^2} - \theta_0^T V_0^{-1}\right]\theta \right\}}$$

Bayesian linear regression

$$\theta_n^T V_n^{-1} - \frac{Y^T X}{\sigma^2} - \theta_0^T V_0^{-1} = 0 \qquad \text{when } \theta_n = V_n \left[V_0^{-1}\theta_0 + \frac{X^T Y}{\sigma^2} \right]$$

and when this happens, we have:

$$P(\theta \mid X, Y, \sigma^2) \propto e^{-\frac{1}{2}(\theta - \theta_n) V_n^{-1} (\theta - \theta_n)}$$

By the definition of a multivariate Gaussian, we have:

$$\int e^{-\frac{1}{2}(\theta-\theta_n)^T V_n^{-1}(\theta-\theta_n)} d\theta = |2\pi V_n|^{1/2}$$

$\therefore$

$$P(\theta \mid X, Y, \sigma^2) = |2\pi V_n|^{-1/2} e^{-\frac{1}{2}(\theta-\theta_n)^T V_n^{-1}(\theta-\theta_n)}$$

Bayesian linear regression

Consider the special case where $\boldsymbol{\theta}_0 = \mathbf{0}$ and $\mathbf{V}_0 = \tau_0^2 \mathbf{I}_d$, which is a spherical Gaussian prior. Then the posterior mean reduces to

$$\begin{aligned} \boldsymbol{\theta}_n &= \frac{1}{\sigma^2}\mathbf{V}_N \mathbf{X}^T \mathbf{y} = \frac{1}{\sigma^2}\left(\frac{1}{\tau_0^2}\mathbf{I}_d + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y} \\ &= \left(\lambda \mathbf{I}_d + \mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y} \end{aligned}$$

where we have defined $\lambda := \frac{\sigma^2}{\tau_0^2}$. We have therefore recovered **ridge regression** again!

Bayesian versus ML plugin prediction

Posterior mean: $\theta_n = \left(\lambda \mathbf{I}_d + \mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$

Posterior variance: $V_n = \sigma^2 \left(\lambda \mathbf{I}_d + \mathbf{X}^T\mathbf{X}\right)^{-1}$

To predict, Bayesians marginalize over the posterior. Let x_* *be a new input. The prediction, given the training data* $D=(X, y)$, *is:*

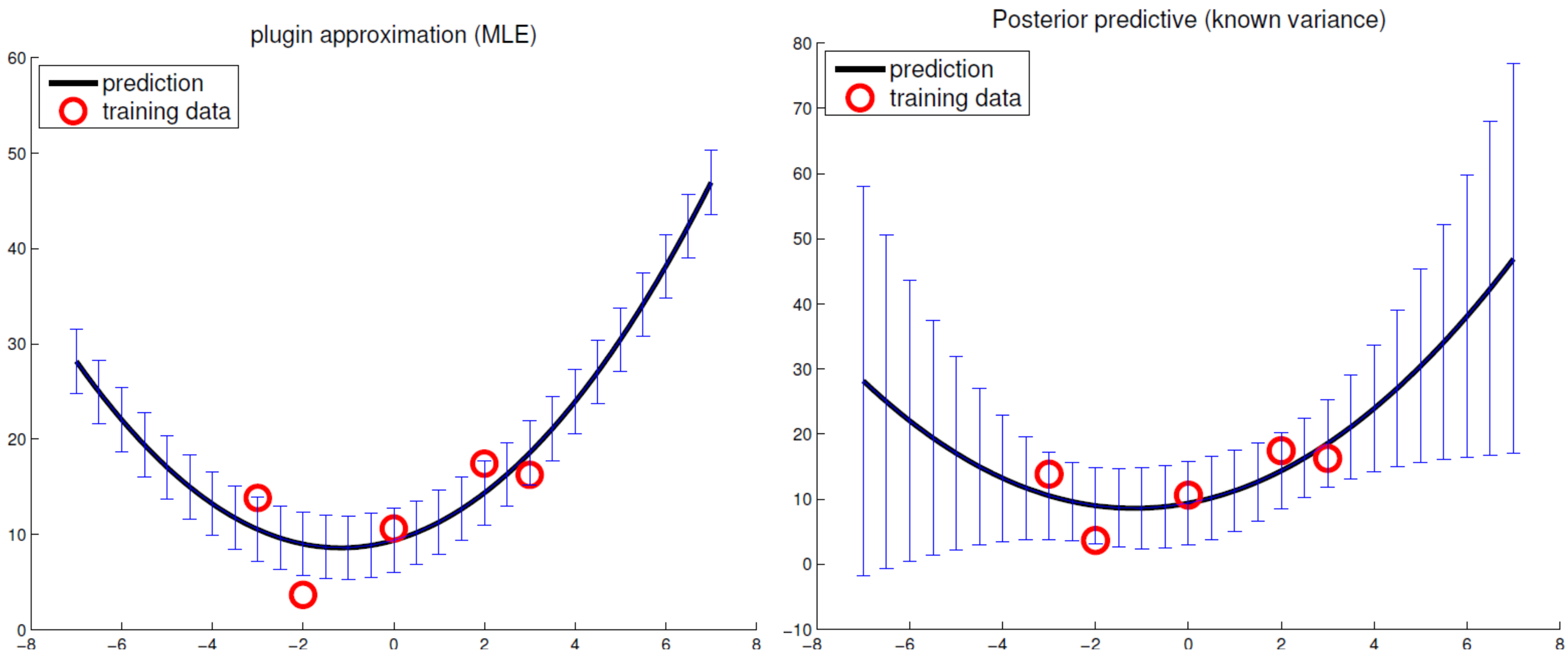
$$P(y|x_*, D, \sigma^2) = \int \mathcal{N}(y|x_*^T\theta, \sigma^2)\, \mathcal{N}(\theta|\theta_n, V_n)\, d\theta$$

$$= \mathcal{N}(y|x_*^T\theta_n, \sigma^2 + x_*^T V_n x_*)$$

On the other hand, the ML plugin predictor is:

$$P(y|x_*, D, \sigma^2) = \mathcal{N}(y|x_*^T\theta_{ML}, \sigma^2)$$

Bayesian versus ML plug-in prediction



Next lecture

In the next lecture, we capitalize on what we have learned for linear models and attack the problem of nonlinear prediction.