IFT6390
Fondements de l’apprentissage machine

Linear regression
and
Regularized linear regression

Professor: Ioannis Mitliagkas
Slides: Pascal Vincent

Supervised task

predict y from x

input $\mathbf{x} \in \mathbb{R}^d$

target (label) y

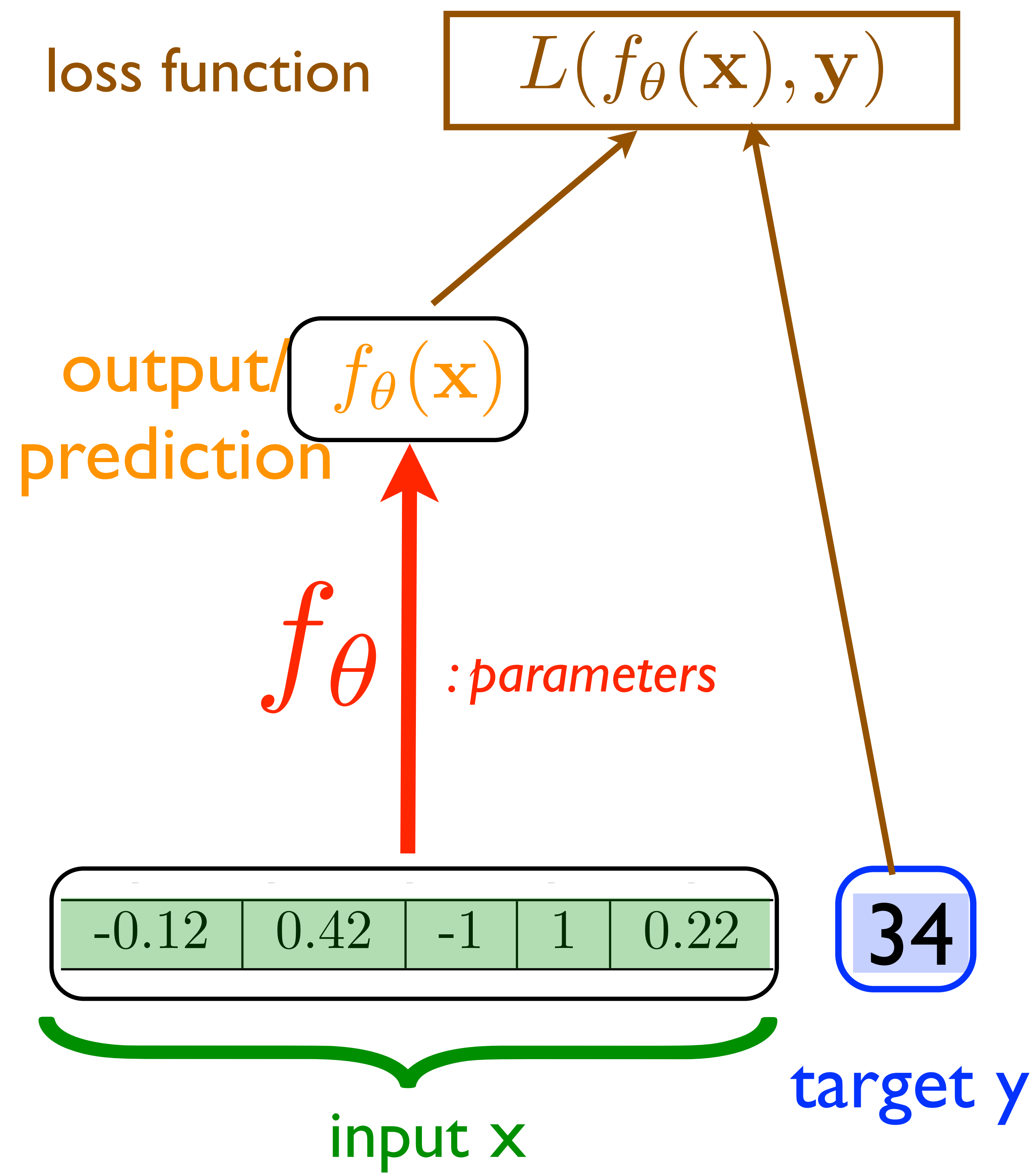
n examples

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	t
0.32	-0.27	+1	0	0.82	113
-0.12	0.42	-1	1	0.22	34
0.06	0.35	-1	1	-0.37	56
0.91	-0.72	+1	0	-0.63	77
...	...	...	...	...	...

Training set Dn

$$D_n = \{(x^{(1)}, t^{(1)}), \ldots, (x^{(n)}, t^{(n)})\}$$

Learn a function $f = f_\theta$ which minimizes a cost (loss).



Empirical risk minimization

We must specify:

- A parametric form for our functions, f_θ
- A specific cost (loss) function $L(y, t)$

So we define the empirical risk as:

$$\hat{R}(f_\theta, D_n) = \sum_{i=1}^{n} L(f_\theta(\mathbf{x}^{(i)}), t^{(i)})$$

i.e. total loss on the training set

Learning amounts to finding the optimal values for the parameters:

$$\theta^\star = \arg\min_\theta \hat{R}(f_\theta, D_n)$$

It is the principle of empirical risk minimization.

Eg: Linear regression

A very simple learning algorithm

We select

A linear (affine) form for the function:

$$f_\theta(\mathbf{x}) = \underbrace{\langle \mathbf{w}, \mathbf{x} \rangle}_{\text{scalar product (inner product)}} + b$$

$$\theta = \{\mathbf{w}, b\}, \underbrace{\mathbf{w}}_{\text{“weight vector”}} \in \mathbb{R}^d, \underbrace{b}_{\text{bias}} \in \mathbb{R}$$

Cost: quadratic error:

$$L(y, t) = (y - t)^2$$

t---> y^

Principle of empirical risk minimization (ERM)

We look for the parameters that minimize the empirical risk

$$\theta^\star = \arg\min_\theta \hat{R}(f_\theta, D_n)$$

Linear regression

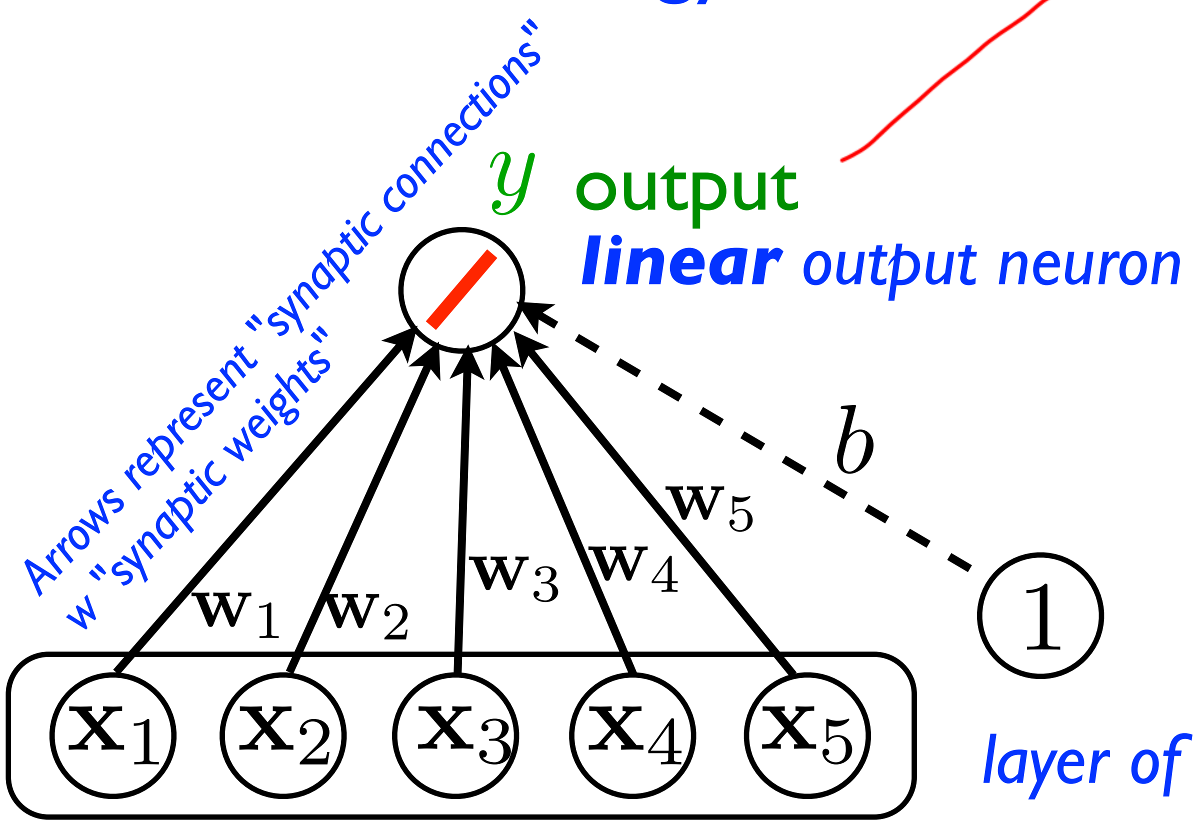
Neural inspiration

Intuitive understanding of the scalar product

each component of x has a weighted influence on the output y

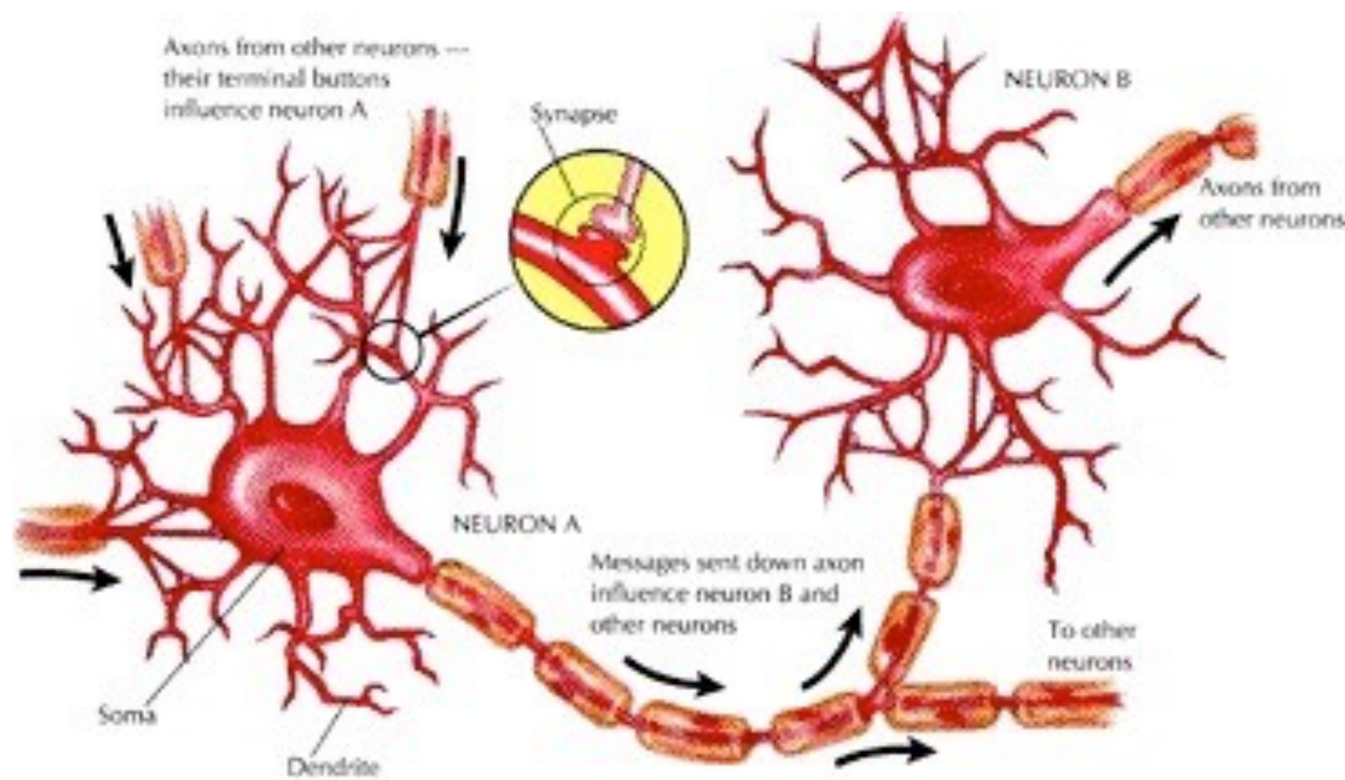
$$\boxed{y = f_\theta(\mathbf{x}) = \mathbf{w}_1\mathbf{x}_1 + \mathbf{w}_2\mathbf{x}_2 + \ldots + \mathbf{w}_d\mathbf{x}_d + b}$$

Neural terminology



layer of input neurons

input (observation) x



Regularized empirical risk

It is often necessary to induce a "preference" for some parameter values rather than others to avoid overfitting

We define regularized empirical risk as follows:

$$\hat{R}_\lambda(f_\theta, D_n) = \underbrace{\left(\sum_{i=1}^{n} L(f_\theta(\mathbf{x}^{(i)}), t^{(i)})\right)}_{\text{empirical risk}} + \underbrace{\lambda\Omega(\theta)}_{\text{regularization term (penalty)}}$$

When we dont have additional data, we can use reg. to avoid overfitting

If we go toward, the complex model more, then the penalize alswo will be more

Ω penalizes more or less the different parameter values.

$\lambda \geq 0$ the importance of this regularization term

(in relation to the empirical risk)

Eg: Ridge Regression

= linear regression + quadratic (L2) regularization

We penalize the large weights

$$\Omega(\theta) = \Omega(\mathbf{w}, b) = \|\mathbf{w}\|^2 = \sum_{j=1}^{d} \mathbf{w}_j^2$$

Neural terminology:
"weight decay" penalty

$$\hat{R}_\lambda(f_\theta, D_n) = \underbrace{\left(\sum_{i=1}^{n} L(f_\theta(\mathbf{x}^{(i)}), t^{(i)})\right)}_{\text{empirical risk}} + \underbrace{\lambda\Omega(\theta)}_{\text{regularization term (penalty)}}$$

Eg: Ridge regression

= linear regression + quadratic (L2) regularization

L2 also known as weight decay(in neural network)

Regularized empirical risk

$$\hat{R}_\lambda(f_\theta, D_n) = \underbrace{\left(\sum_{i=1}^{n} L(f_\theta(\mathbf{x}^{(i)}), t^{(i)})\right)}_{\text{Empirical risk}} + \underbrace{\lambda\Omega(\theta)}_{\substack{\text{regularization term} \\ \text{(penalty)}}}$$

We are looking for the parameter values that minimize this objective

$$\{\mathbf{w}^*, b^*\} = \theta^* = \underset{\theta}{\operatorname{argmin}} \hat{R}_\lambda(f_\theta, D_n)$$

Eg: Ridge Regression

= linear regression + quadratic (L2) regularization

- For linear regression or ridge regression a little linear algebra gives us an **analytical solution**

we solve for $\theta = \{b, \mathbf{w}\}$: $$\frac{\partial \hat{R}_\lambda(f_\theta, D_n)}{\partial \theta} = 0$$

we obtain: $$\boxed{\begin{pmatrix} b^* \\ \mathbf{w}^* \end{pmatrix} = (\check{X}^T \check{X} + \lambda \check{I})^{-1} \check{X}^T \mathbf{t}}$$

This is a closed form solution

$$\text{où } \check{X} = \begin{pmatrix} 1 & \mathbf{x}_1^{(1)} & \dots & \mathbf{x}_d^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \mathbf{x}_1^{(n)} & \dots & \mathbf{x}_d^{(n)} \end{pmatrix}, \mathbf{t} = \begin{pmatrix} t^{(1)} \\ \vdots \\ t^{(n)} \end{pmatrix} \quad \check{I} = \begin{pmatrix} 0 & 0 & & & \\ 0 & 1 & 0 & & \\ & 0 & 1 & 0 & \\ & & 0 & \ddots & 0 \\ & & & 0 & 1 \end{pmatrix}$$

- Most of the time (other choices of f and L) we do not have an analytical solution.
- More generally, we can use a gradient descent method.

$f(x) = x^T\theta \qquad \theta \equiv w$, no bias

$D_n = \{(x_1, y_1), \dots, (x_n, y_n)\} \qquad x_i \in \mathbb{R}^d$

$y_i \in \mathbb{R}$

$X = [x_1, x_2, \dots x_n]^T \in \mathbb{R}^{n \times d}$

$y = [y_1, y_2 \dots y_n]^T \in \mathbb{R}^n$

Least squares regression (no ridge)

$\lambda = 0$

$L(y, t) = (y-t)^2$

$\hat{R}(\theta, D_n) = \sum_{i=1}^{n} (y_i - \langle x_i, \theta \rangle)^2 = \|y - X\theta\|_2^2$

Prove that $\theta^* = (X^TX)^{-1}X^Ty$ * Assumption

$J(\theta) = \|y - X\theta\|_2^2 = (y - X\theta)^T(y - X\theta)$

$J(\theta) = \|y - X\theta\|_2^2 = (y - X\theta)^T(y - X\theta)$ **

$\nabla_\theta J = -2X^T(y - X\theta)$

Generalization of scalar rule:

$g(\theta) = f(h(\theta))$

$g'(\theta) = (h(\theta))' f'(h(\theta))$

$J(\theta) = (y - X\theta)^T(y - X\theta) =$

$= y^Ty - y^TX\theta - \theta^TX^Ty + \theta^TX^TX\theta$

$\nabla_\theta J = -2X^T(y - X\theta) \overset{\text{set}}{=} 0$

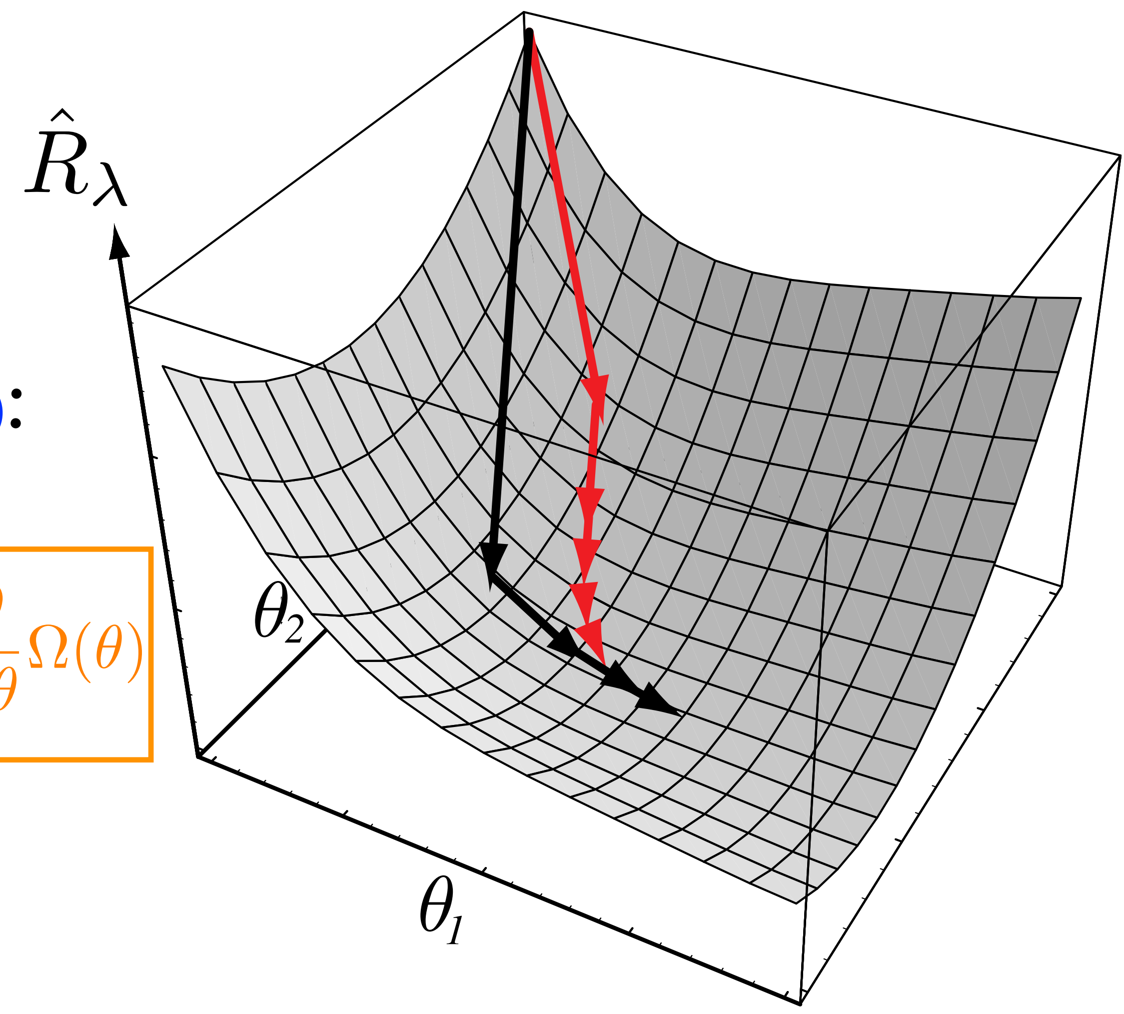
$X^TX\theta^* = X^Ty$

$\theta^* = (X^TX)^{-1}X^Ty$

Other possibility:
optimization by gradient descent

$$\hat{R}_\lambda(f_\theta, D_n) = \underbrace{\left(\sum_{i=1}^{n} L(f_\theta(\mathbf{x}^{(i)}), t^{(i)})\right)}_{\text{empirical risk}} + \underbrace{\lambda\Omega(\theta)}_{\text{regularization term}}$$

- we initialize the parameters randomly
- we update them iteratively following the gradient



Either batch gradient descent (whole dataset):

Loop: $\theta \leftarrow \theta - \eta \frac{\partial \hat{R}_\lambda}{\partial \theta}$

$$= \boxed{\left(\sum_{i=1}^{n} \frac{\partial}{\partial \theta} L(f_\theta(\mathbf{x}^{(i)}), t^{(i)})\right) + \lambda \frac{\partial}{\partial \theta}\Omega(\theta)}$$

Or stochastic gradient descent:

Loop:

For i in 1...n

$$\theta \leftarrow \theta - \eta \frac{\partial}{\partial \theta}\left(L(f_\theta(\mathbf{x}^{(i)}), t^{(i)}) + \frac{\lambda}{n}\Omega(\theta)\right)$$

Or other variants of the gradient descent idea

(conjugate gradient, Newton's method, natural gradient, ...)

Various regularizers

«Ridge»: regularization, L_2

In Bayesian terms: corresponds to a Gaussian prior on the weights

$$\Omega(\theta) = \Omega(\mathbf{w}, b) = \|\mathbf{w}\|_2^2 = \sum_{j=1}^{d} \mathbf{w}_j^2$$

«Lasso»: regularization, L_1

In Bayesian terms: corresponds to a Laplacian prior on the weights

$$\Omega(\theta) = \Omega(\mathbf{w}, b) = \|\mathbf{w}\|_1 = \sum_{j=1}^{d} |\mathbf{w}_j|$$

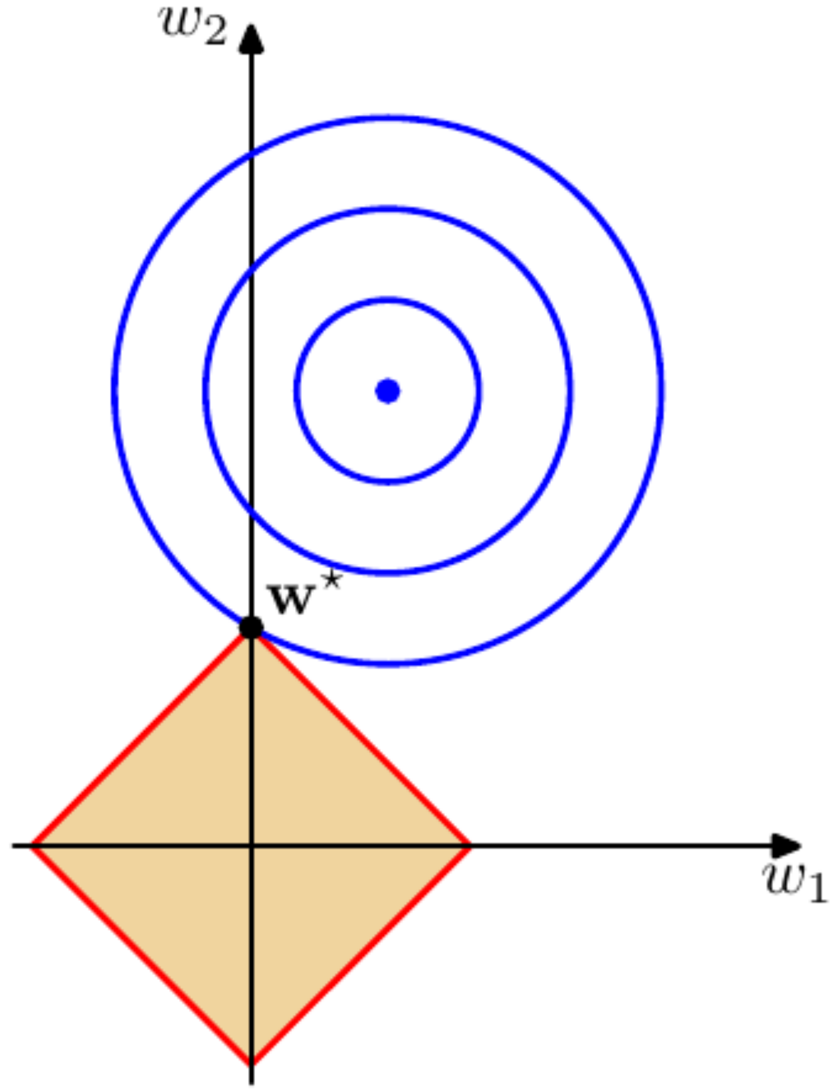
=> automatic selection of components (a number of weights will be zero)

«Elastic net»: combination of the two

$$\Omega(\theta) = \Omega(\mathbf{w}, b) = \lambda_1 \|\mathbf{w}\|_1 + \lambda_2 \|\mathbf{w}\|_2^2$$

Etc...

Visualizing L_1 regularization

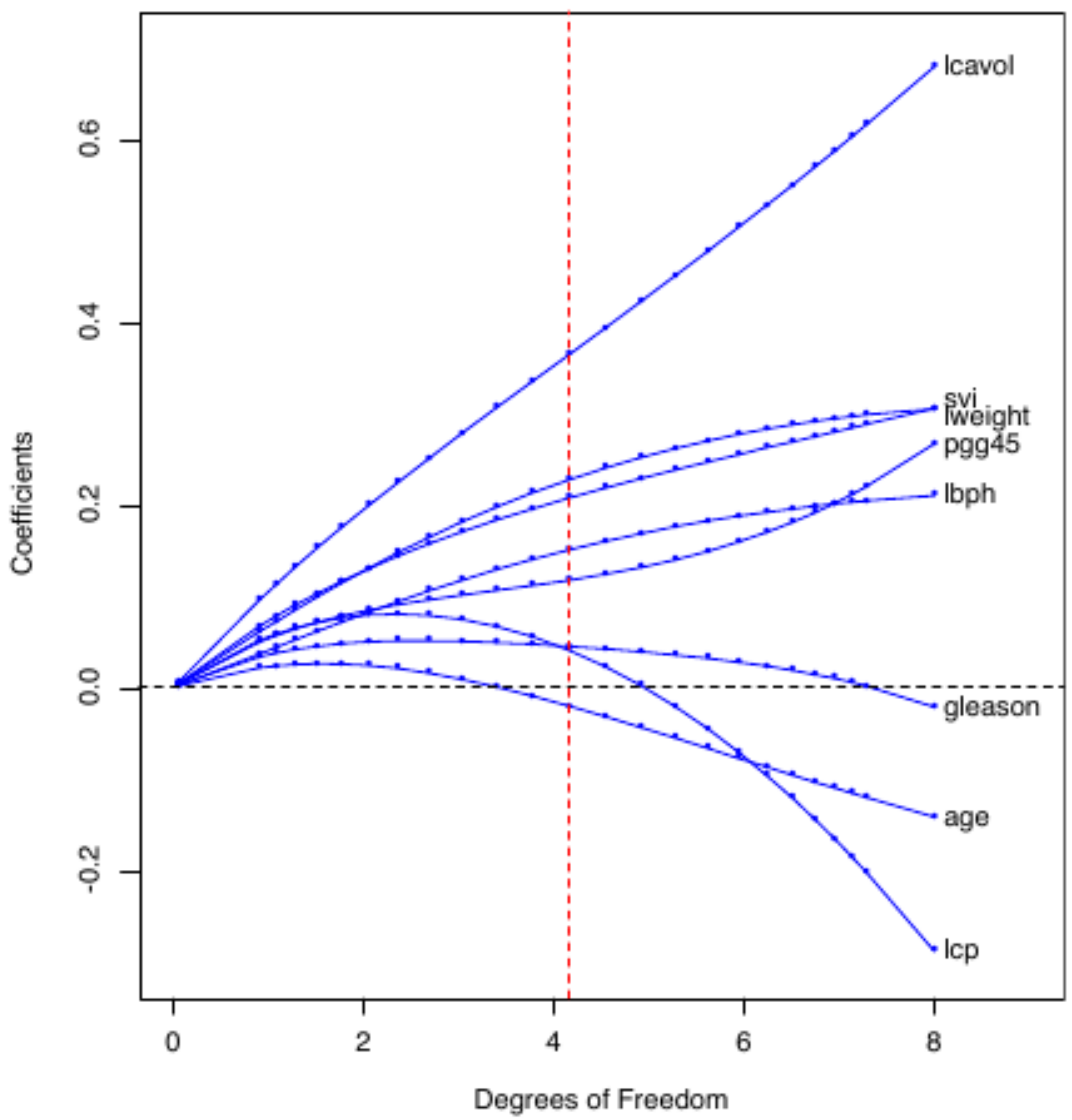


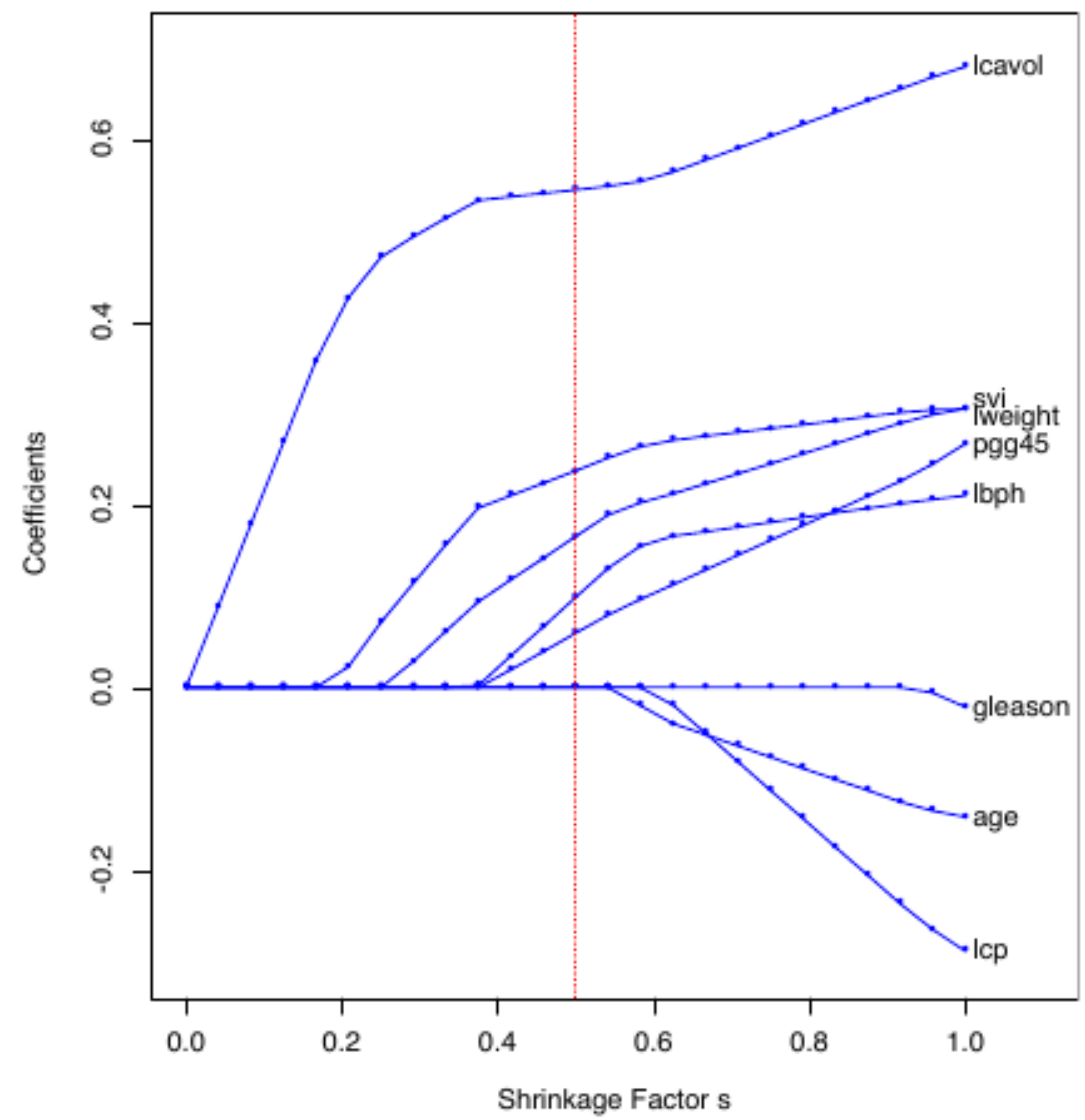
- If λ is big enough, the circle is very likely to intersect the diamond at one of the corners
- This makes L_1 regularization much more likely to make some weights *exactly* 0

Pros and cons of L_1 regularization

- If there are irrelevant input features, Lasso is likely to make their weights 0, while L_2 is likely to just make all weights small
- Lasso is biased towards providing *sparse solutions* in general
- Lasso optimization is computationally more expensive than L_2
- More efficient solution methods have to be used for large numbers of inputs (e.g. least-angle regression, 2003).
- L_1 methods of various types are very popular
- One can combine L_1 and L_2 regularization (elastic-net)

Example of L1 vs L2 effect





- Note the sparsity in the coefficients induces by L_1
- Lasso is an efficient way of performing the L_1 optimization