Regularization & Sparsity

Marco Loog

IN4320 MACHINE LEARNING

TU Delft
Delft University of Technology

Outline

- Data & generalization →
  Empirical risk →
  Regression →
  Stability →
  Regularization →
  Sparsity →
  Time for more Qs?

  - Q : indicates a Q you should be able to answer…

TU Delft

The Setting

- Say we have N feature vectors x_i and corresponding outputs or targets y_i
- Say we want to estimate a functional relationship $f(x; w) \approx y$, with parameters w, to predict correct outputs to new and unseen feature vectors
- Q : how could we do this?

TU Delft

Empirical Risk

- All we have is N observations, so we could try and find a w that at least works well on these
- "Working well" is expressed in terms of loss ℓ
- Total loss on all points is the empirical risk

$$\sum_{i=1}^{N} \ell(f(x_i; w), y_i)$$

TU Delft

Minimizing Empirical Risk

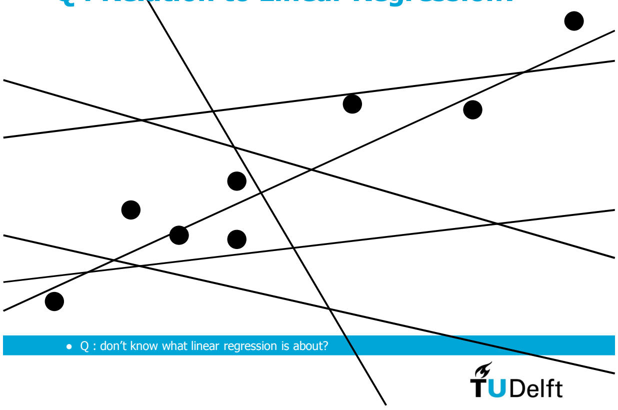
- One often considers the solution with the minimal empirical risk

$$\underset{w}{\operatorname{argmin}} \sum_{i=1}^{N} \ell(f(x_i; w), y_i)$$

  - Q : approaches you know that fit formulation?

TU Delft

Q : Relation to Linear Regression?



- Q : don't know what linear regression is about?

TU Delft

Linear Least Squares Regression

- Standard regression solves

$$\min_w \sum_{i=1}^{N} (f(x_i, w) - y_i)^2$$

- Where f is linear in x : $f(x; w) = w^T x$

TUDelft

Linear Least Squares Regression

- Solution

$$w = (XX^T)^{-1} XY^T$$

X matrix with all x_i in columns; Y output row

- In case of too few observations, we need pseudo-inverse : $w = (XX^T)^{+} XY^T$

- Understand how to come to these solutions!!!

TUDelft

Many Dimensions / Few Observations

- What happens with relatively few observations in relatively high dimensions?

- E.g. assume average x_i is 0 and consider $w = (XX^T)^{-1} XY^T = \left(\frac{1}{N} XX^T\right)^{-1} \left(\frac{1}{N} XY^T\right)$

  - Q : eigenvalues of the covariance matrix?
  - Q : effect of this on the vector XY^T?
  - Do experiments if you do not see or believe…

TUDelft

Many Dimensions / Few Observations

- Solution $w = (XX^T)^{-1} XY^T$ is unstable and can be all over the place
- Generalization to unseen data can, and will often, be very bad

- Q : how to stabilize the solution? Any ideas?

TUDelft

Stabilization, One Way to Perform

- Idea : keep eigenvalues away from 0

- Add identity to XX^T : $w = (XX^T + \lambda I)^{-1} XY^T$

- Q : why consider the identity?

TUDelft

Stabilization as Regularization

- Idea : keep eigenvalues away from 0

- Add identity to XX^T : $w = (XX^T + \lambda I)^{-1} XY^T$

- This choice of w is, in fact, the solution of

$$\min_w \sum_{i=1}^{N} (f(x_i, w) - y_i)^2 + \lambda \|w\|^2$$

TUDelft

An Equivalent View

- Instead of solving

$$\min_w \sum_{i=1}^{N} (f(x_i, w) - y_i)^2 + \lambda \|w\|^2$$

one can also solve

$$\min_w \sum_{i=1}^{N} (f(x_i, w) - y_i)^2$$

$$\text{s.t. } \|w\|^2 \leq \tau$$

TUDelft

Short Intermezzo?

- What is the shape of these functions?

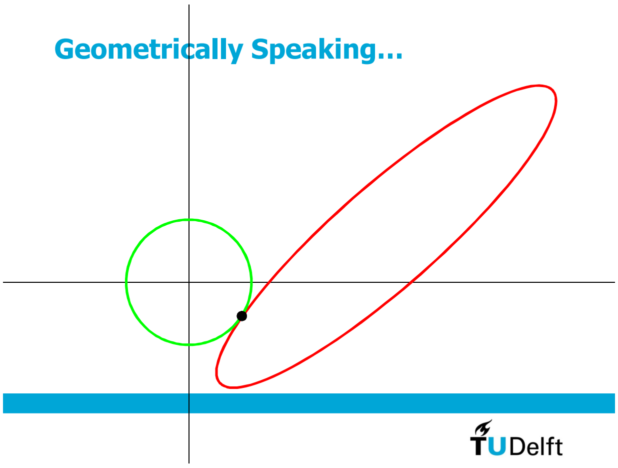
$$\sum_{i=1}^{N} (f(x_i, w) - y_i)^2 + \lambda \|w\|^2$$

$$\sum_{i=1}^{N} (f(x_i, w) - y_i)^2$$

$$\|w\|^2$$

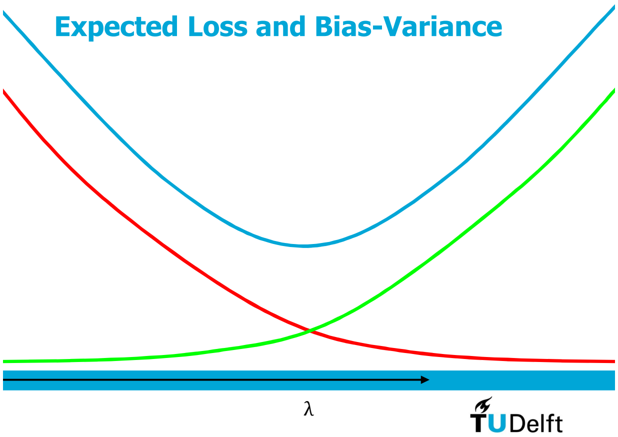
TUDelft

Geometrically Speaking…



TUDelft

Expected Loss and Bias-Variance



TUDelft

Regularized Risk

- General approach to regularization

$$\min_w \sum_{i=1}^{N} \ell(f(x_i, w), y_i) + R(f)$$

- Many learning problems in PR and ML can be [and are in fact] formulated in this way
- Different considerations give different R
- Various links : MAP, MDL, SRM, etc.

TUDelft

Introducing Sparsity

- For a change, let us consider

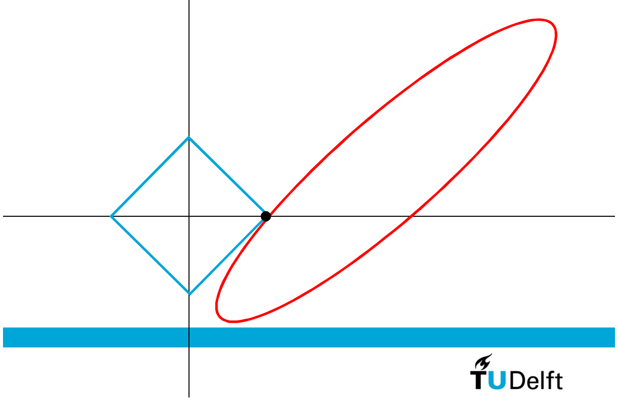
$$\min_w \sum_{i=1}^{N} (f(x_i, w) - y_i)^2$$

$$\text{s.t. } \|w\|_1 \leq \tau$$

- Q : what is the shape of $\|w\|_1$?
- Q : what is the effect of this change of norm?

TUDelft

The Geometry…



Again the Equivalent View…

- Include sparsifying norm as an additive term

$$\min_{w} \sum_{i=1}^{N} (f(x_i, w) - y_i)^2 + \lambda \|w\|_1$$

- Matlab "demo"…?

Final Remarks

- Sparsity by regularization due to Tibshirani
  - Least absolute shrinkage and selection operator or lasso
  - Performs feature selection
  - Compare to feature forward selection etc.!
- Regularization framework also used for classification…

?