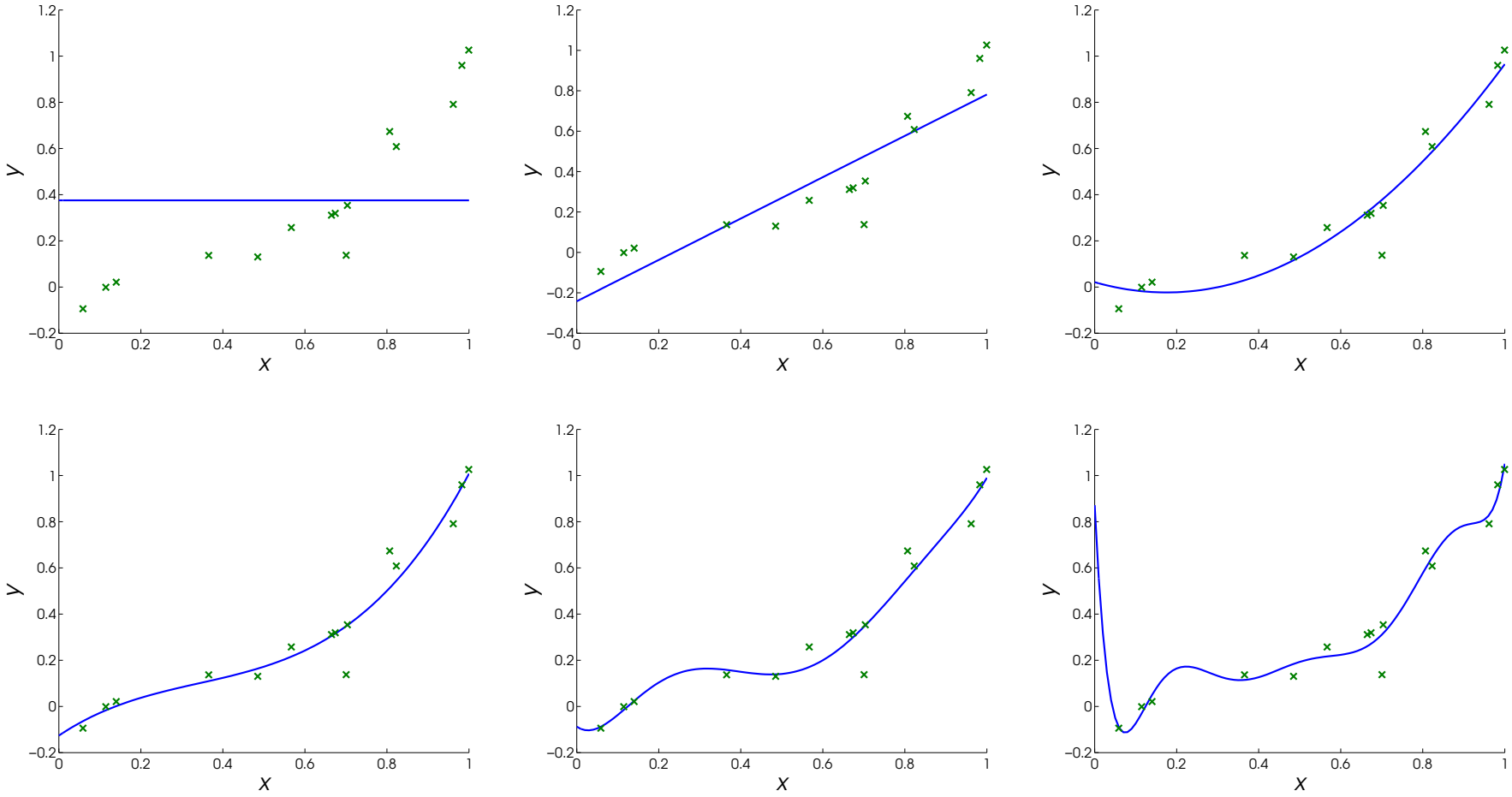
Regularization and Cross Validation

Fernando Pérez Cruz

25 de October 2004

Polynomial Approximations



SOLUTION

- We used the following approximations:

$$f(y) = \prod_{i=1}^{P} w_i x^i + w_0$$

- Weight vector and training error using square-loss:

model	w_0	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	error
Degree 0	0.38	0	0	0	0	0	0	0	0	0.1176
Degree 1	-0.24	1.02	0	0	0	0	0	0	0	0.0220
Degree 2	0.02	-0.51	1.45	0	0	0	0	0	0	0.0074
Degree 3	-0.13	1.19	-2.32	2.26	0	0	0	0	0	0.0059
Degree 6	-0.09	-1.26	27.5	-119	215	-174	52.8	0	0	0.0052
Degree 8	0.87	-36.1	474	2860	9349	17615	19101	11056	2644	0.0036

- We get little improvement for $P \geq 2$ and the weights are much larger.

REGULARIZATION

- Minimizing the loss for flexible models $\Rightarrow$ **Overfitting**.
- From the above example, we know that the weights do not need to grow indefinitely to get a good approximation.
- Regularization Theory:

$$R_{reg}(\mathbf{w}) = R_{emp}(\mathbf{w}) + \nu\Omega(\mathbf{w})$$

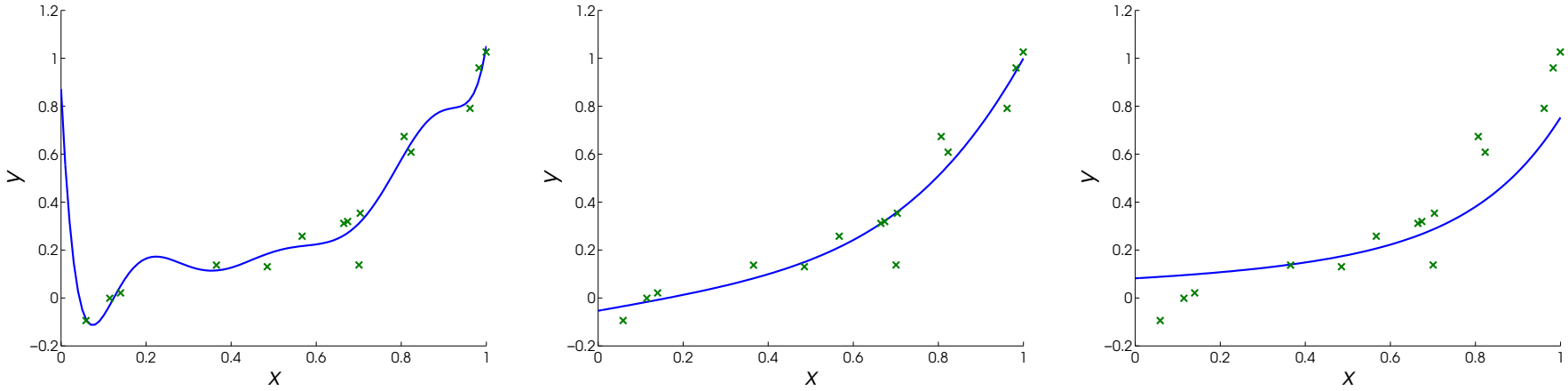
- $R_{emp}(\mathbf{w})$ is the empirical risk.
- $\Omega(\mathbf{w}) \geq 0$ regulariser.
- ν trade-off hyperparameter.

Weight Decay

$$\Omega(\mathbf{w}) = \|\mathbf{w}\|^2$$

- It prefers lower weights.
- If a weight increase does not significantly reduce $R_{emp}(\mathbf{w})$, then $\Omega(\mathbf{w})$ will keep it low.
- It is also known as:
  - Ridge Regression
  - Logarithm of Gaussian Prior (MAP)
  - Maximum margin (SVMs)
  - It is related to: early stopping and training with noise.

EXAMPLE



ν	w_0	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8
0	0.87	-36.1	474	2860	9349	17615	19101	11056	2644
0.01	-0.053	0.312	0.053	0.156	0.241	0.231	0.147	0.023	-0.110
10	0.082	0.104	0.098	0.091	0.084	0.079	0.075	0.072	0.069

ν	0	0.01	10
Tr. Error	0.0036	0.0062	0.0243
Error	0.0271	0.0092	0.0226

HYPERPARAMETER SELECTION

- We have seen models for nonlinear regression:

$$\min_{\mathbf{w}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \nu \|\mathbf{w}\|^2.$$

$$f(x) = \sum_{j=1}^{P} w_j x_j^P + w_0$$

- We have seen Neural Networks for classification:

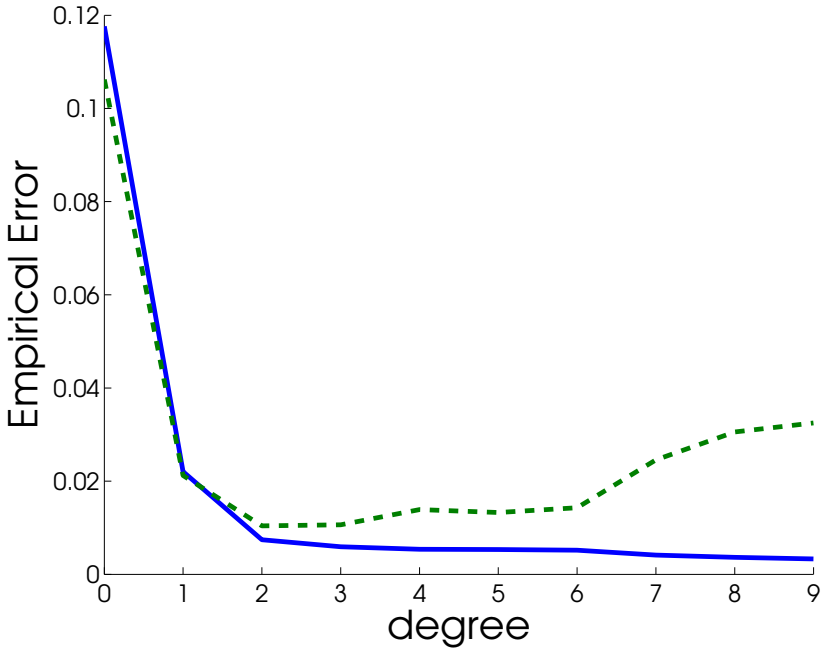
$$\min_{\mathbf{w},\mathbf{v}_j} \frac{1}{2n} \sum_{i=1}^{n} (y_i \log(f(\mathbf{x}_i)) + (1 - y_i) \log(1 - f(\mathbf{x}_i)))^2 + \nu \|\mathbf{w}\|^2.$$

$$f(\mathbf{x}) = \sum_{j=1}^{N} w_j \frac{1}{1 + e^{-\mathbf{v}_j^\top \mathbf{x} - v_{j0}}} + w_0$$

- But how do we choose ν, P and N?

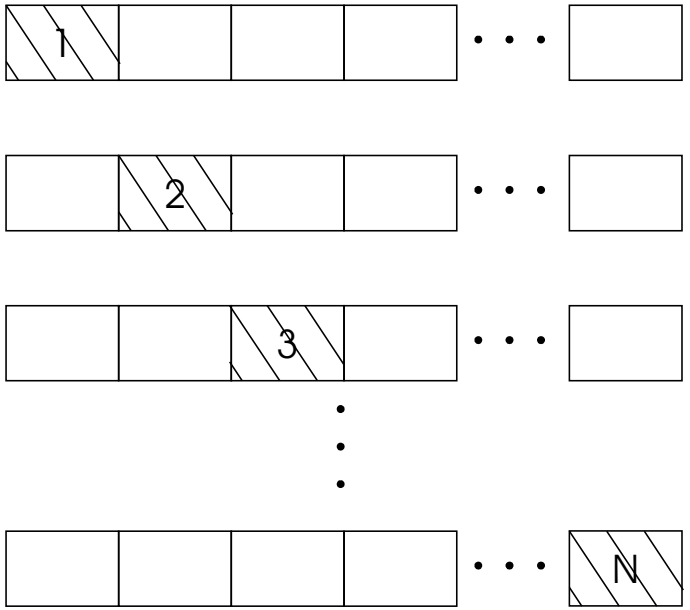
VALIDATION SET

- Cannot rely on the training error $\Rightarrow$ the more versatile $f(\mathbf{x})$ the lower it will be.
- Empirical error over a validation set.
- For small datasets we will be "wasting" samples not using them for training.
- Moreover, the final error needs to be obtained from other set (test set), otherwise the solution will be biased by the validation set.



Cross-Validation

- Divide the training set in N subsets.
- Leave the first subset, train the "machine" with the remaining sets.
- Compute empirical error for the 1^{st} set.
- Repeat N times, leaving each time one of the subset for validation.
- Select the hyperparameters using the smallest mean validation error.



- Typical $N = 10$. $N = 1$, known as leave-one-out error.