Empirical Risk Minimization

Expected and Empirical Error

Given a function f, a loss function V, and a probability distribution μ over Z = X x Y, the **expected or true error of f is:**

$$I[f] = \int_z V\big(y, f(x)\big) d\mu(z)$$

In general μ is unknown. Thus, given n data points the empirical error of f is

$$I_s[f] = \frac{1}{n} \sum V(y_i, f(x_i))$$

ERM

Given a training set S and a function space $\mathcal{H}$, empirical risk minimization (Vapnik introduced the term) is the class of algorithms that look at S and select f_S as

$$f_S = \arg\min_{f \in H} I_S[f]$$

For example in linear regression ERM, loss function is V(z) = $(f(x)-y)^2$ and $\mathcal{H}$ is space of linear functions f = ax.

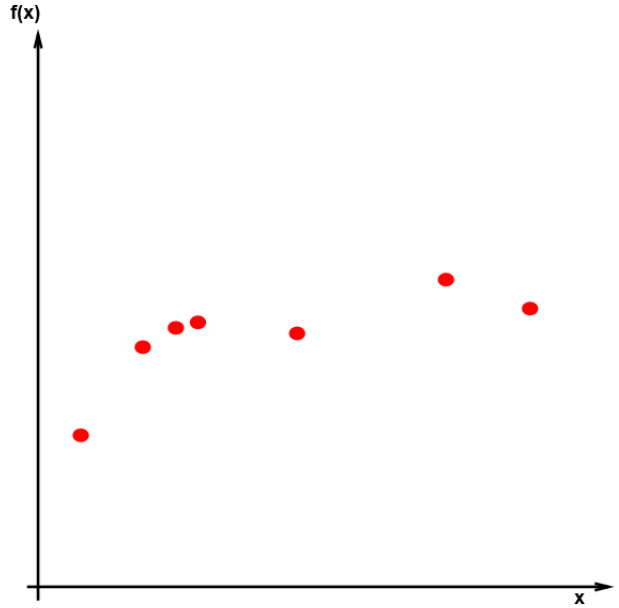
Generalization and Well-posedness of Empirical Risk Minimization

For ERM to represent a “good” class of learning algorithms, the solution should

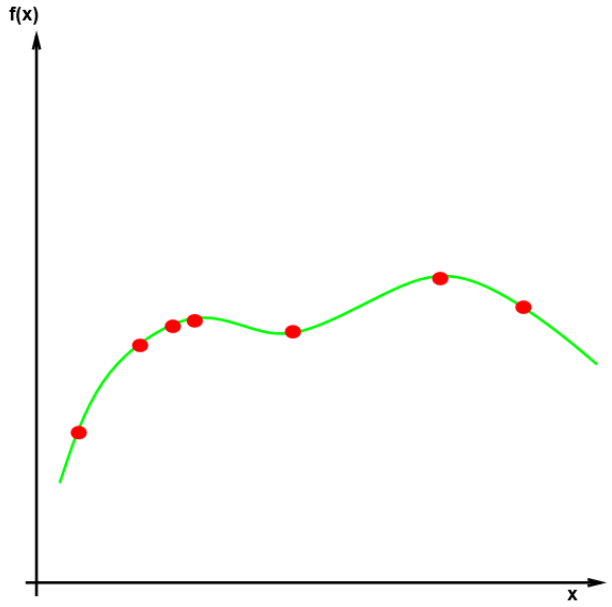
- Generalize - how accurately an algorithm is able to predict outcome values for previously unseen data.

- Exist, be unique and – especially – be stable (well-posedness).

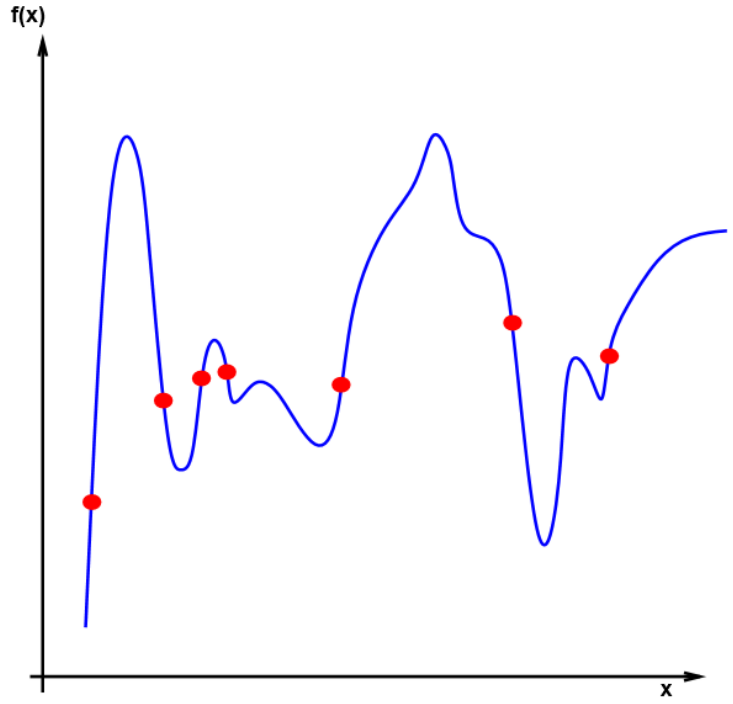
ERM and generalization: given a certain number of samples…



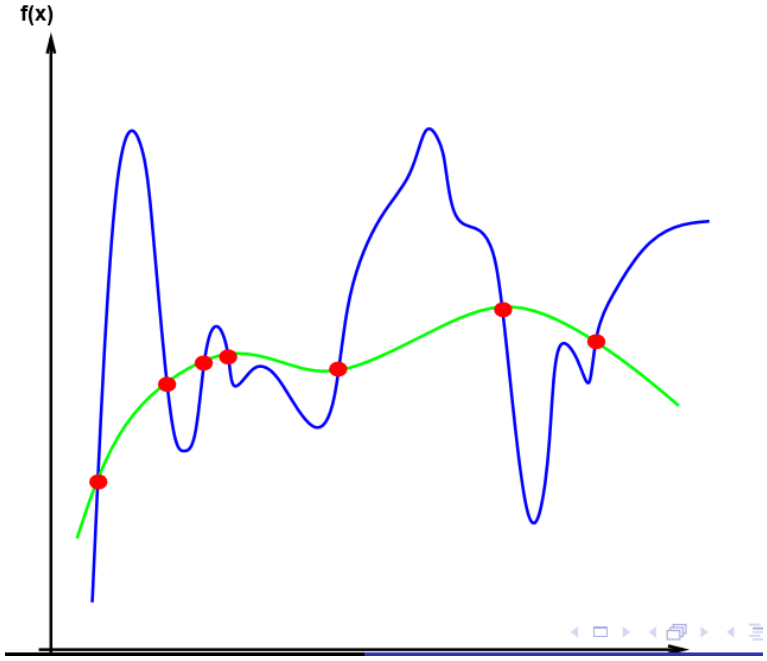
…suppose this is the “true” solution…



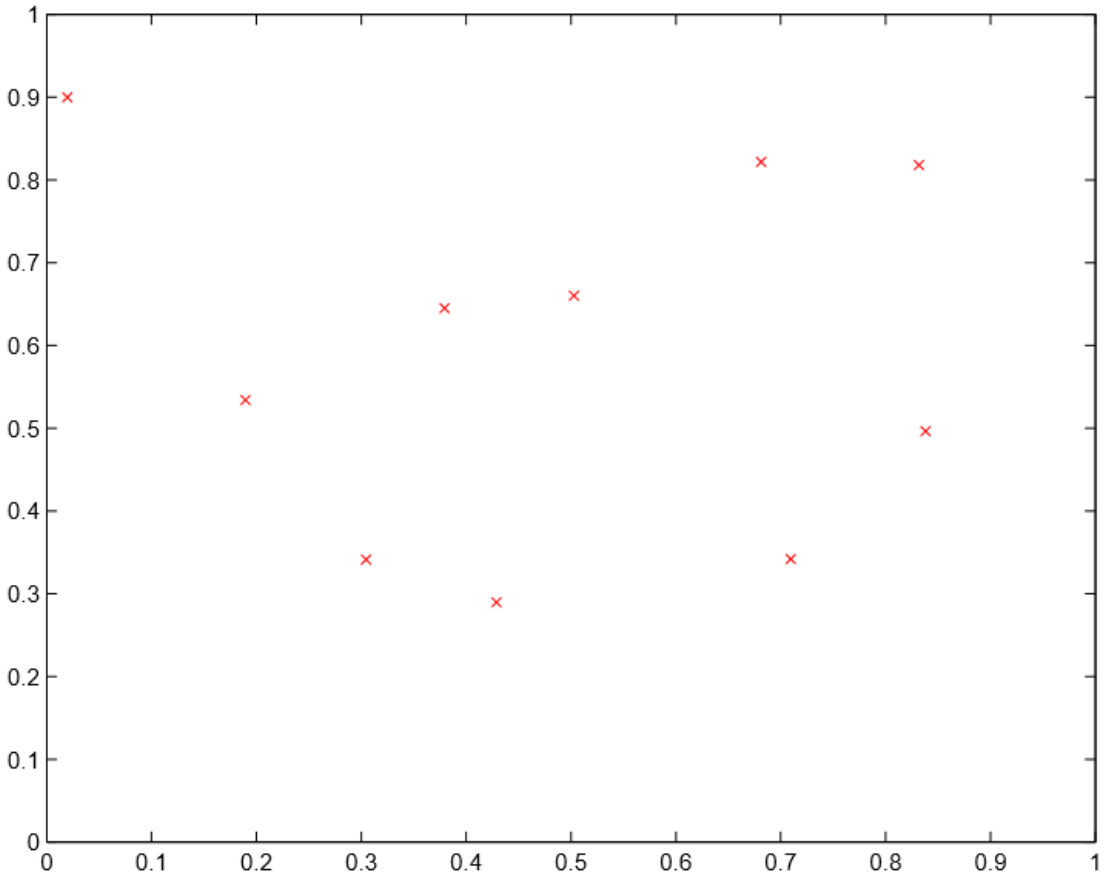
… but suppose ERM gives this solution.



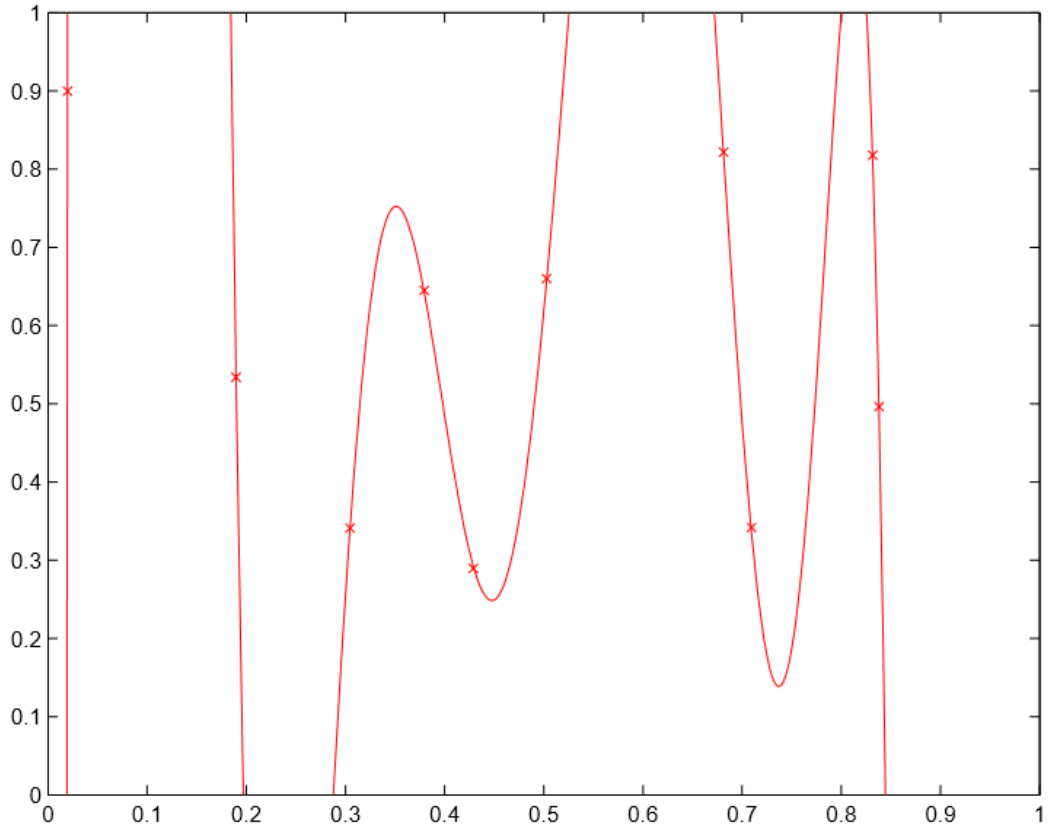
ERM solution is not generalized: Solution does not accurately model the true outcomes



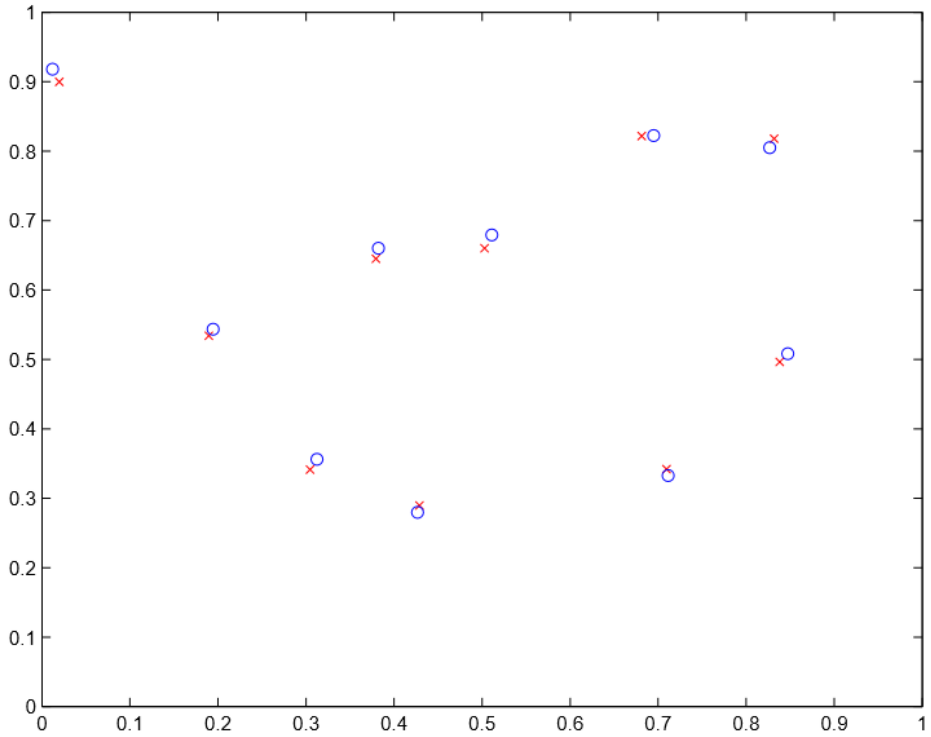
ERM and stability: given 10 samples…



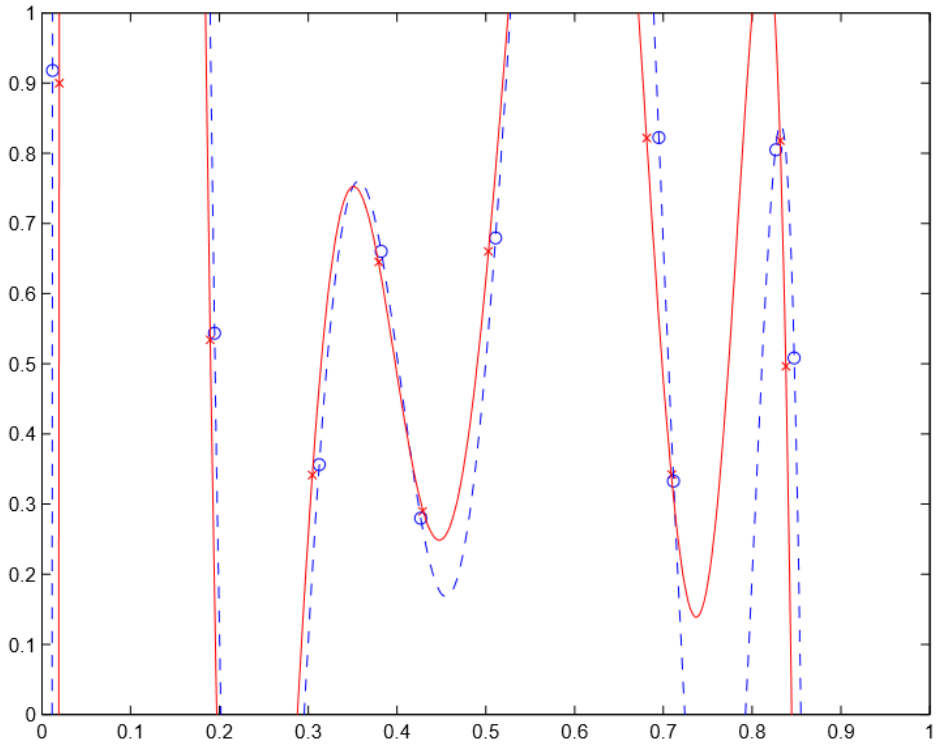
...we can find the smoothest interpolating polynomial (which degree?).



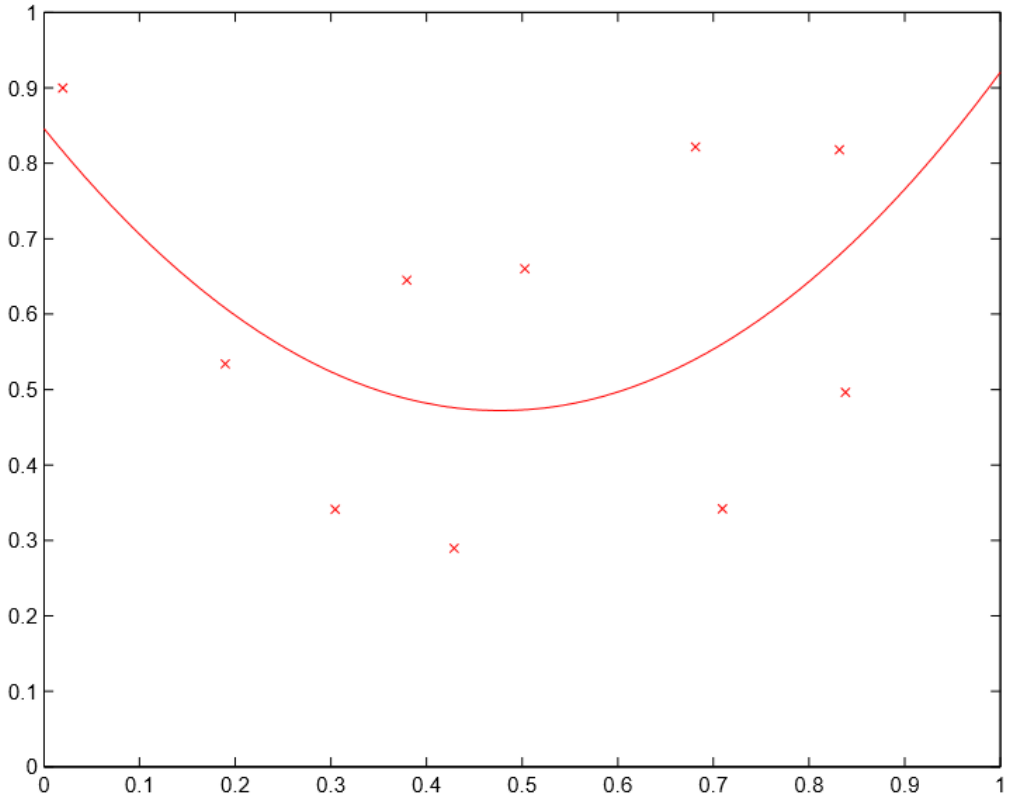
But if we perturb the points slightly…



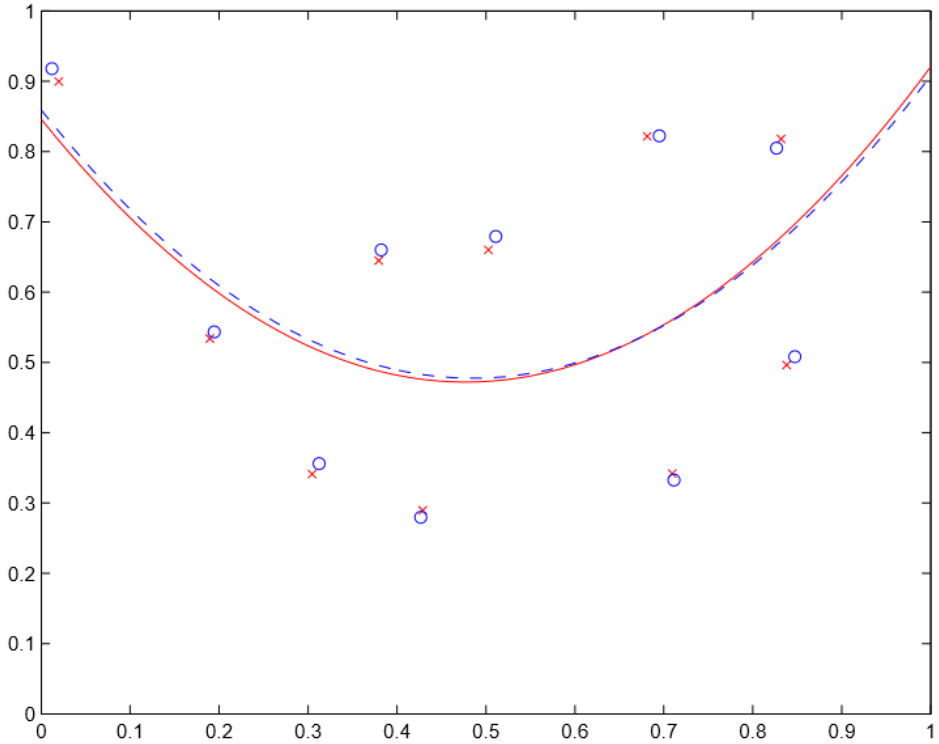
...the solution changes a lot!



If we restrict ourselves to degree two polynomials…



...the solution varies only a small amount under a small perturbation.



Regularization

Regularized least squares (RLS) is a family of methods for solving the least-squares problem while using regularization to further constrain the resulting solution.

Partial list of RLS methods [edit]

The following is a list of possible choices of the regularization function $R(\cdot)$, along with the name for each one, the corresponding prior if there is a simple one, and ways for computing the solution to the resulting optimization problem.

Name	Regularization function	Corresponding prior	Methods for solving
Lasso regression	$\|w\|_1$	Laplace	Proximal gradient descent, least angle regression
Tikhonov regularization	$\|w\|_2^2$	Normal	Closed form
ℓ_0 penalization	$\|w\|_0$	–	Forward selection, Backward elimination, use of priors such as spike and slab
Elastic nets	$\beta\|w\|_1 + (1-\beta)\|w\|_2^2$	–	Proximal gradient descent
Total variation regularization	$\sum_{j=1}^{d-1} \|w_{j+1} - w_j\|$	–	Split-Bregman method, among others

Regularization

General form

$$\min_{f \in \mathcal{H}} \left[\sum V(y_i, f(x_i)) + \lambda J(f) \right]$$

Where J(f) is penalty functional and $\mathcal{H}$ is the space of functions on which J(f) is defined

Tikhonov Regularization

- Tikhonov regularization ensures well-posedness. eg existence, uniqueness and especially stability of the solution

- Tikhonov regularization ensures generalization - how accurately the outcome values for previously unseen data are predicted.

Tikhonov Regularization

Suppose the problem is represented as

$$Xf = Y$$

Ordinary LS minimizes

$$\|Y - Xf\|^2$$

In this case, the solution is $\hat{f} = (X^T X)^{-1} X^T Y$

In Tikhonov Regularization, we minimize

$$\|Y - Xf\|^2 + \|\Gamma f\|^2$$

Where, Γ is the Tikhonov matrix. In this case the solution is given as $\hat{f} = (X^T X + \Gamma^T \Gamma)^{-1} X^T Y$

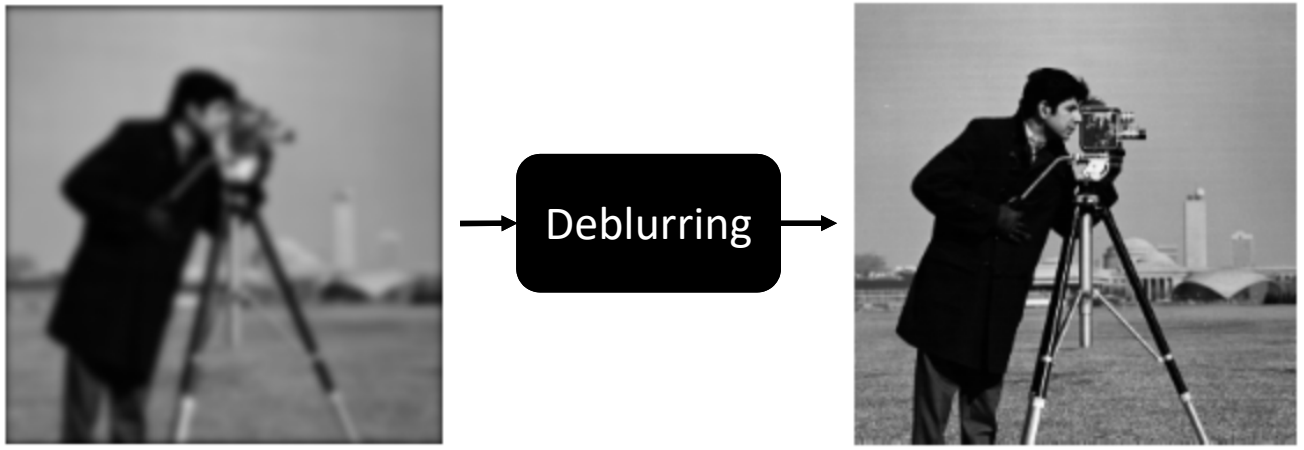
$$\hat{\beta}^{ridge} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \boldsymbol{y}, \qquad (3.44)$$

Image Restoration - Example

The problem is to find the true (original) image, I, given:

- the blurred image, J
- Blurring operator, H

such that, $J = HI + \epsilon$, where ϵ is the noise in data.



Blurred Image, J

Restored Image, I

Image Restoration - Example

Generalized least square solution (OLS) is given as:

- $\hat{I} = H^{-1}J$
- $\hat{I} = (H^T H)^{-1} H^T J$

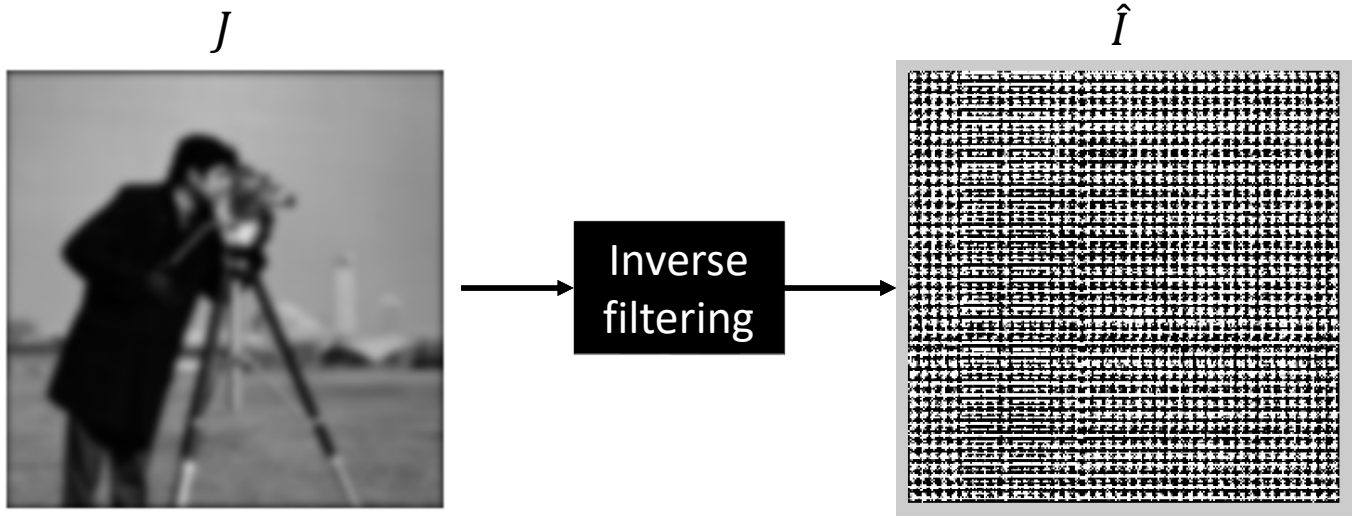


Image Restoration - Example

Using Tikhonov Regularization:

- $\hat{I} = (H^T H + \alpha^2 L^T L)^{-1} H^T J$, where $\Gamma = \alpha L$

Early Stopping

Early Stopping

- To regularize non-parametric regression problem encountered in machine learning.
- **Nonparametric regression** is a category of regression analysis in which the predictor does not take a predetermined form but is constructed according to information derived from the data.
- One common choice of regression function is to use functions from a Reproducing Kernel Hilbert Space (RKHS).
- RKHS is infinite dimensional, in which they can supply solutions that overfit the training sets of arbitrary size.
- In these cases, regularization is important.
- One way to regularize non-parametric regression problems is to apply an early stopping rule to an iterative procedure such as gradient descent.

Gradient Descent

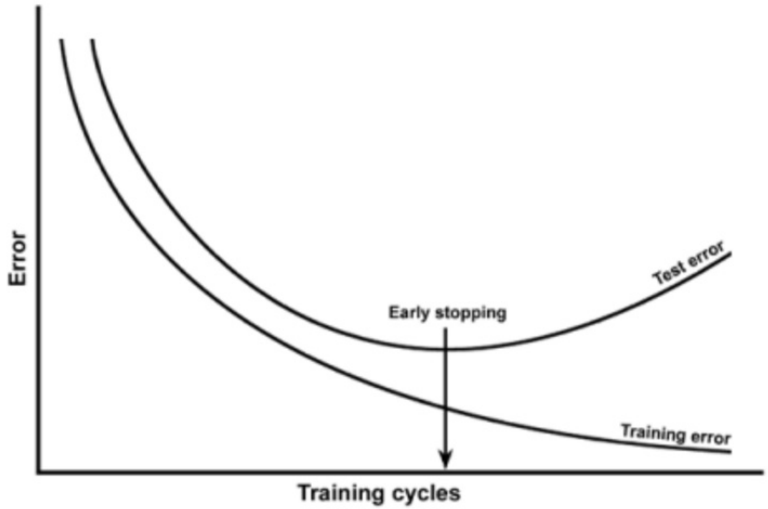
The components of f are updated in the direction of the negative gradient:

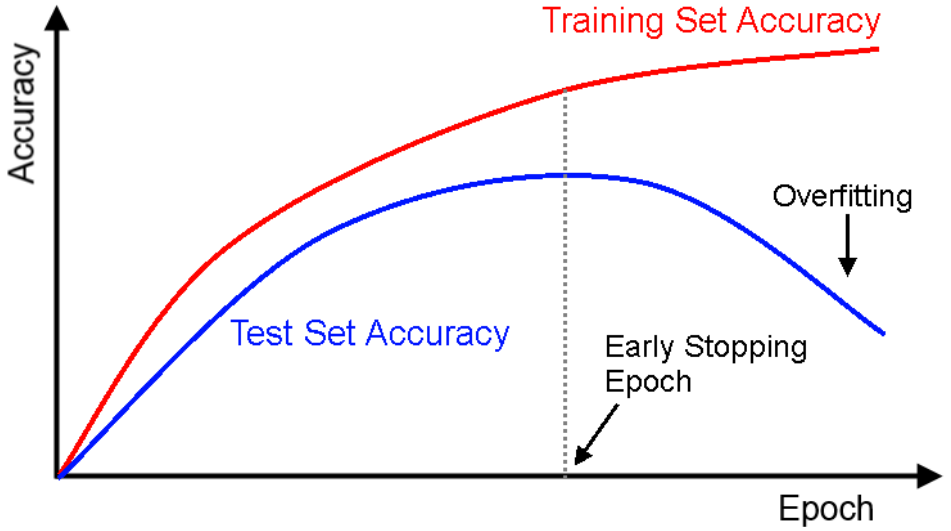
$$f^{\tau+1} = f^{\tau} - \delta \nabla E(f^{\tau})$$

where τ is the iteration number, δ is the step size and $E(.)$ is an error function

Early stopping rule are based on the analysis of upper bounds of the generalization error as a function of iteration number.

Example plots showing performance of Early Stopping





Steps in Validation-based Early Stopping

1. Split the training data into a training set and a validation set, e.g. in a 2-to-1 proportion.
2. Train only on the training set and evaluate the per-example error on the validation set once in a while, e.g. after every fifth epoch.
3. Stop training as soon as the error on the validation set is higher than it was the last time it was checked.
4. Use the weights the network had in that previous step as the result of the training run.

Structural Risk Minimization

SRM

- ERM suggests to minimize the empirical risk at any cost

- SRM looks for the optimal relationship between
  - the amount of empirical data
  - the quality of approximation of the data by the function chosen from a given set of functions
  - the value that characterizes capacity of a set of functions

Result 6.1 (Vapnik's Book)

With probability at least $1-\eta$ simultaneously for all functions from the set of totally bounded functions $0 \leq Q(z,\alpha) \leq B, \alpha \in \Lambda$, with finite VC dimension h the following (additive) inequality holds true:

$$R(\alpha) \leq R_{emp}(\alpha) + \frac{B\varepsilon(l)}{2}\left(1+\sqrt{1+\frac{4R_{emp}(\alpha)}{B\varepsilon(l)}}\right)$$

Where

$R(\alpha) = \int Q(z,\alpha)dF(z)\,, z \in Z$, is the risk function

$R_{emp}(\alpha)$ is the empirical risk function

$\varepsilon(l) = 4\frac{h\left(\ln\frac{2l}{h}+1\right)-\ln\frac{\eta}{4}}{l}$, l is the number of training samples

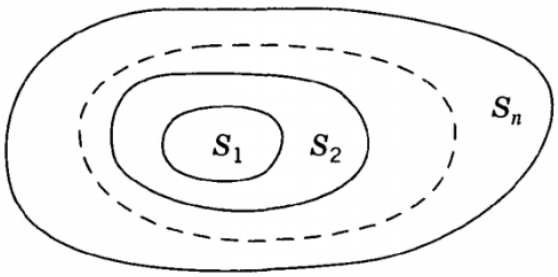
SRM

- If it happens that $\frac{l}{h}$ is large, then the value of actual risk is determined by the value of empirical risk. Therefore to minimize actual risk one minimizes the empirical risk.
- However, if $\frac{l}{h}$ is small, a small value of empirical risk $R_{emp}(\alpha)$ does not guarantee a small value of the actual risk.
- The first term in inequality depends on a specific function of the set of functions, while for a fixed number of observations the second term depends mainly on the VC dimension of the whole set of functions.
- Thus, to minimize both terms simultaneously, VC dimension is made controlling variable.

Imposing Structure

- Impose the structure $\mathcal{S}$ on the set S of functions $Q(z, \alpha), \alpha \in \Lambda$
- The set of nested subsets of functions:

$$S_1 \subset S_2 \subset \cdots \subset S_m$$



where, $S_k = \{Q(z, \alpha): \alpha \in \Lambda_k\}$ and

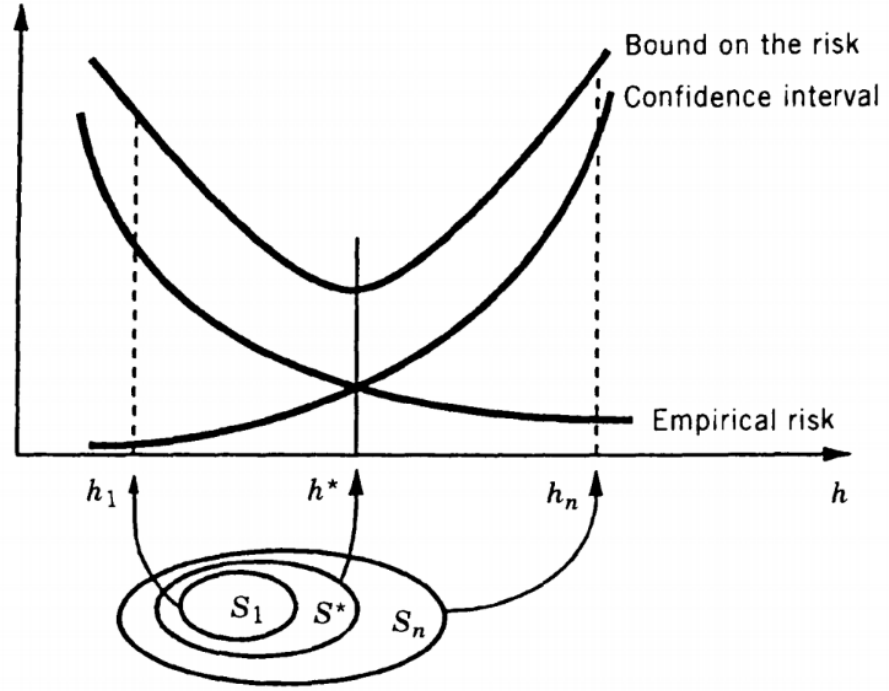
$$S = \bigcup_k S_k$$

- The sequence of values of VC dimensions h_k for the elements S_k of the structure $\mathcal{S}$ is non-decreasing with increasing k

$$h_1 \leq h_2 \leq \cdots \leq h_m$$

SRM Induction Principle

"To provide the given set of functions with an admissible structure and then to find the function that minimizes guaranteed risk over given elements of the structure."



Steps in SRM

1. Using a priori knowledge of the domain, choose a class of functions , such as polynomials of degree n , neural networks having n hidden layer neurons, a set of splines with n nodes or fuzzy logic models having n rules .
2. Divide the class of functions into a hierarchy of nested subsets in order of increasing complexity. For example, polynomials of increasing degree.
3. Perform empirical risk minimization on each subset (this is essentially parameter selection).
4. Select the model in the series whose sum of empirical risk and VC confidence is minimal.

References

1. V. N. Vapnik, “Statistical Learning Theory”. Wiley, 1998.
2. T. Hastie, R.Tibshirani, J. Friedman, “The Elements of Statistical Learning:Data Mining, Inference and Prediction”, Springer Series in Statistics, 2009
3. Al Bovik, “Handbook of Image & Video Processing”, Elsevier Academic Press, 2005
4. Tomaso Poggio, “The Learning Problem and Regularization”, Lecture Notes, 2010.
5. “Early Stopping” - Wikipedia