Model Selection and Regularization Parameter Choice

Regularization Methods for High Dimensional Learning

Francesca Odone and **Lorenzo Rosasco**
odone@disi.unige.it - lrosasco@mit.edu

July 2, 2014

About this class

Goal To discuss the choice of the regularization parameter, giving a brief description of the theoretical results and an overview of a few heuristics used in practice.

Plan

- The general problem: model selection
- Error analysis: sketch
  - error decomposition: the bias variance trade-off
  - sample error
  - approximation error
- Heuristics

Regularization parameter

We have seen that a learning algorithm can be seen as a map

$$S \to f_S$$

from the training set to the hypotheses space.

Actually most learning algorithms define a one parameter family of solutions, i.e.

given $\lambda > 0$

$$S \to f_S^\lambda$$

Examples

- Tikhonov regularization
- Spectral regularization
- Sparsity based regularization
- Manifold regularization

but also SVM, boosting....

In all these algorithms one (or more parameters) control the smoothness of the solution and has to be tuned to find the final solution.

Actual algorithms can often be seen as a two step procedures.

Optimal choice

We can start asking:

1. whether there exists an optimal parameter choice
2. what it depends on
3. and most importantly if we can design a scheme to find it.

Remember that our goal is to have *good generalization properties*...

Oracle choice

Recall that the training set S is sampled i.i.d with respect to $p(x, y)$.

Oracle choice

Recall that the training set S is sampled i.i.d with respect to $p(x, y)$.
Ideally we want to choose λ to make the generalization error *small*.

Oracle choice

Recall that the training set S is sampled i.i.d with respect to $p(x, y)$.
Ideally we want to choose λ to make the generalization error *small*.
Recall that

$$I[f] = \int_{X \times Y} V(f(x), y) p(x, y) dx dy$$

Oracle choice

Recall that the training set S is sampled i.i.d with respect to $p(x, y)$.
Ideally we want to choose λ to make the generalization error *small*.
Recall that

$$I[f] = \int_{X \times Y} V(f(x), y) p(x, y) dxdy$$

Excess Risk

$$I[f_S^\lambda] - \inf_f \{ I[f]$$

Consistency

A minimal requirement on the parameter choice $\lambda = \lambda_n$ is that it should lead to consistency.

Consistency

For all $\epsilon > 0$

$$\lim_{n \to \infty} P\{ I[f_S^{\lambda_n}] - \inf_f I[f] \} > \epsilon\} = 0$$

Probabilistic Bound

A possible approach is that of:
1) finding a suitable probabilistic bound for fixed λ,
2) minimizing the bound w.r.t. λ.

Probabilistic Bound

A possible approach is that of:
1) finding a suitable probabilistic bound for fixed λ,
2) minimizing the bound w.r.t. λ.

For $\lambda > 0$, and for $0 < \eta \leq 1$, with probability at least $1 - \eta$

$$I[f_S^\lambda] - \inf_f \{I[f]\} \leq \epsilon(\eta, \lambda, n)$$

A PRIORI PARAMETER CHOICE

We can then define the parameter choice $\lambda^* = \lambda(\eta, n)$ minimizing the bound, i.e.

$$\min_{\lambda} \epsilon(\eta, \lambda, n)$$

We have yet to see how to find a bound...

Error Decomposition

The first step is, often, to consider a suitable error decomposition

$$I[f_S^\lambda] - \inf_f\{I[f]\} = I[f_S^\lambda] - I[f^\lambda] + I[f^\lambda] - \inf_f\{I[f]\}$$

The function f^λ is the *infinite sample* regularized solution, for example in Tikhonov regularization is the solution of

$$\min_{f\in\mathcal{H}} \int_{X\times Y} V(f(x), y)p(x, y)dxdy + \lambda\|f\|_{\mathcal{H}}^2$$

Error Decomposition (cont.)

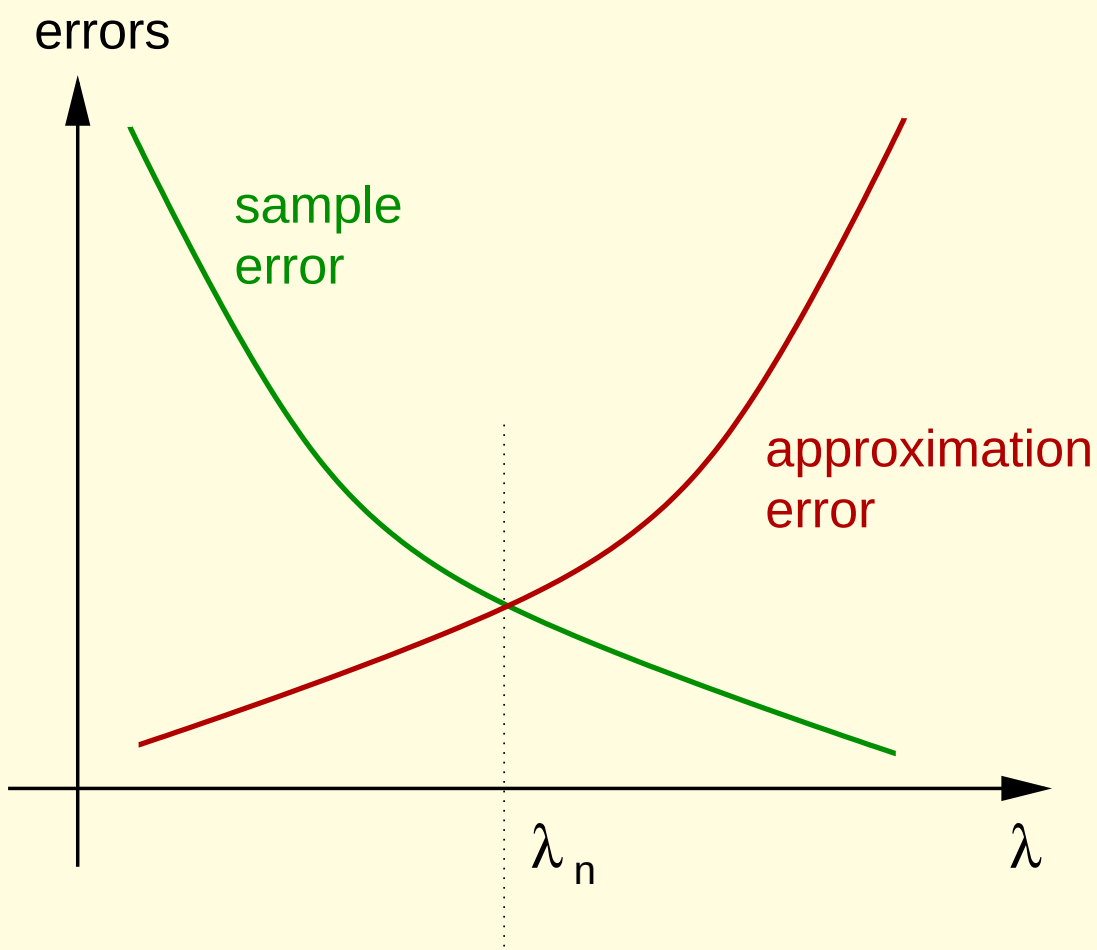
Consider

$$I[f_S^\lambda] - \inf_f\{I[f]\} = \underbrace{I[f_S^\lambda] - I[f^\lambda]}_{\text{sample error}} + \underbrace{I[f^\lambda] - \inf_f\{I[f]\}}_{\text{approximation error}}$$

- The sample error $I[f_S^\lambda] - I[f^\lambda]$ quantifies the error due to finite sampling/noise
- The approximation error $I[f^\lambda] - \inf_f\{I[f]\}$ quantifies the bias error due to the chosen regularization scheme

TRADE-OFF

The two terms typically have the following behavior



The parameter choice λ^* solves a bias variance trade-off.

Sample Error

$$\underbrace{I[f_S^\lambda] - I[f^\lambda]}_{\text{sample error}}$$

To study the sample error, we have to compare empirical quantities to expected quantities.
The main mathematical tools we can use are quantitative version of the law of large numbers.

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} V(f(x_i), y_i) + \lambda \|f\|_{\mathcal{H}}^2$$

$$\min_{f \in \mathcal{H}} \int_{X \times Y} V(f(x), y) p(x, y) dx dy + \lambda \|f\|_{\mathcal{H}}^2$$

If $\xi_1, \ldots, \xi_n$ i.i.d. zero mean real random variables and $|\xi_i| \leq C, \quad i = 1, \ldots, n$, then Höeffding inequality ensures that $\forall \epsilon > 0$

$$P\{|\frac{1}{n}\sum_i \xi_i| \geq \epsilon\} \leq 2e^{-\frac{n\epsilon^2}{2C^2}}$$

or equivalently setting $\tau = \frac{n\epsilon^2}{2C^2}$ we have with probability at least (with confidence) $1 - 2e^{-\tau}$

$$|\frac{1}{n}\sum_i \xi_i| \leq \frac{C\sqrt{2\tau}}{\sqrt{n}}.$$

The case of Tikhonov Regularization

The explicit form of the sample error is typically of the form

$$I[f_S^\lambda] - I[f^\lambda] \leq \frac{C\sqrt{\log \frac{2}{\eta}}}{\lambda n}$$

where the above bound holds with with probability at least $1 - \eta$.

If λ decreases sufficiently slow the sample error goes to zero as n increases.

Approximation error

Compare f^{λ} and f^{*} solving:

$$\min_{f \in \mathcal{H}} \int_{X \times Y} V(f(x), y) p(x, y) dx dy + \lambda \|f\|^2$$

and

$$\min_{f \in \mathcal{H}} \int_{X \times Y} V(f(x), y) p(x, y) dx dy$$

Approximation error

Compare f^λ and f^* solving:

$$\min_{f \in \mathcal{H}} \int_{X \times Y} V(f(x), y) p(x, y) dx dy + \lambda \|f\|^2$$

and

$$\min_{f \in \mathcal{H}} \int_{X \times Y} V(f(x), y) p(x, y) dx dy$$

- The approximation error is purely deterministic.
- It is typically easy to prove that it decreases as λ decreases.
- The explicit behavior depends on the problem at hand.

The last problem is an instance of the so called *no free lunch theorem*

Few Basic questions

- Can we learn consistently any problem? YES! Consider

$$I[f_S^\lambda] - \inf_f \{I[f]\} \leq \frac{C\sqrt{\log \frac{2}{\eta}}}{\lambda n} + \left(I[f^\lambda] - \inf_f \{I[f]\} \right)$$

with probability at least $1 - \eta$.

Few Basic questions

- Can we learn consistently any problem? YES! Consider

$$I[f_S^\lambda] - \inf_f \{I[f]\} \leq \frac{C\sqrt{\log \frac{2}{\eta}}}{\lambda n} + \left(I[f^\lambda] - \inf_f \{I[f]\} \right)$$

  with probability at least $1 - \eta$.
- Can we always learn at some prescribed speed? NO! it depends on the problem!

The latter statement is the called no free lunch theorem.

Approximation error (cont.)

We have to restrict to a class of problems. Typical examples:

- the target function f^* minimizing $I[f]$ belongs to a RKHS
- the target function f^* belongs to some Sobolev Space with smoothness s
- the target function f^* depends only on a few variables
- ...

Usually the regularity of the target function is summarized in a regularity index r and the approximation error depends on such index

$$I[f^\lambda] - I[f^*] \leq C\lambda^{2r}$$

Putting all together

Putting all together we get with probability at least $1-\eta$,

$$I[f_S^\lambda] - \inf_f\{I[f]\} \leq \frac{C\sqrt{\log\frac{2}{\eta}}}{\lambda n} + C\lambda^{2r}$$

- We choose $\lambda_n = n^{-\frac{1}{2r+1}}$ to optimize the bound
- If we set $f_S = f_S^{\lambda_n}$ we get with high probability

$$I[f_S] - \inf_f\{I[f]\} \leq Cn^{-\frac{2r}{2r+1}}\sqrt{\log\frac{2}{\eta}}$$

Probably Approximatively Correct (PAC)

If we set $f_S = f_S^{\lambda_n}$ we get with high probability

$$I[f_S] - \inf_f \{I[f]\} \leq Cn^{-\frac{2r}{2r+1}} \sqrt{\log \frac{2}{\eta}}$$

There are also other ways of writing the bound.

Sample Complexity (PAC)

If $n \geq n(\epsilon, \eta, r)$, then, with probability at least $1 - \eta$,

$$I[f_S] - \inf_f \{I[f]\} \leq \epsilon$$

Adaptivity

- The parameter choice $\lambda_n = n^{-\frac{2r}{2r+1}}$ depends on the regularity index r which is typically unknown.
- Ideally we would want (adaptive) parameter choices not depending on r and still achieving the rate $n^{-\frac{2r}{2r+1}}$.

The Theory and the Truth

- The bounds are often asymptotically tight and in fact the rates are optimal in a suitable sense

The Theory and the Truth

- The bounds are often asymptotically tight and in fact the rates are optimal in a suitable sense
- Nonetheless we often pay the price of the great generality under which they hold in that they are often pessimistic

The Theory and the Truth

- The bounds are often asymptotically tight and in fact the rates are optimal in a suitable sense
- Nonetheless we often pay the price of the great generality under which they hold in that they are often pessimistic
- Constants are likely to be non optimal

The Theory and the Truth

- The bounds are often asymptotically tight and in fact the rates are optimal in a suitable sense
- Nonetheless we often pay the price of the great generality under which they hold in that they are often pessimistic
- Constants are likely to be non optimal
- In practice we have to resort to heuristics to choose the regularization parameter

The Theory and the Truth

- The bounds are often asymptotically tight and in fact the rates are optimal in a suitable sense
- Nonetheless we often pay the price of the great generality under which they hold in that they are often pessimistic
- Constants are likely to be non optimal
- In practice we have to resort to heuristics to choose the regularization parameter

Why to care about bounds?

typically in proving them we learn something about the problems and the algorithms we use

Hold-out Estimate

One of the most common heuristics is probably the following:

Hold Out Estimates

Split training set S in S_1 and S_2.

- Find estimators on S_1 for different λ.
- Choose λ minimizing the empirical error on S_2.

Repeat for different splits and average answers.

Cross validation

When the data are few other strategies are preferable that are variations of hold-out.

K-fold cross-validation

Split training set S in k groups.

- For fixed λ, train on all $k-1$ groups, test on the group left out and sum up the errors
- Repeat for different values of λ and choose the one minimizing the cross validation error

One can repeat for different splits and average answers to decrease variance

When the data are *really* few we can take $k = n$, this strategy is called leave one-out (LOO)

Implementation: Regularization path

- The real computational price is often the one for finding the solutions for several regularization parameter values (so called regularization path).
- In this view we can somewhat reconsider the different computational prices of spectral regularization schemes.

LOO Implementation for RLS

In the case of Tikhonov regularization there is a simple closed form for the LOO

$$LOO(\lambda) = \sum_{i=1}^{n} \frac{(y_i - f_S^{\lambda}(x_i))}{(I - K(K + \lambda nI)^{-1})_{ii}}$$

If we can compute $(I - K(K + \lambda nI)^{-1})_{ii}$ easily the price of LOO (for fixed λ) is that of calculating f_S^{λ}

It turns out that computing the eigen-decomposition of the kernel matrix can be actually convenient in this case

How Should we Explore the Parameter Range?

In practice we have to choose several things.

- Minimum and maximum value? We can look at the maximum and minimum eigenvalues of the kernel matrix to have a reasonable range (iterative methods don't need any though)
- Step size? the answer is different depending on the algorithm
  - for iterative and projections methods the regularization is intrinsically discrete, for Tikhonov regularization we can take a geometric series $\lambda_i = \lambda_0 q^i$, for some $q > 1$.

What about Kernel Parameters?

So far we only talked about λ and not about kernel parameters

- Both parameters controls the complexity of the solution
- Clearly we can minimize the cross validation error w.r.t. both parameters
- Often a rough choice of a kernel parameter can be allowed if we eventually fine tune λ

For the gaussian kernel a reasonable value of the width σ can be often chosen looking at some statistics of the distances among input points

What about Kernel Parameters?

So far we only talked about λ and not about kernel parameters

- Optimal parameter choice can be defined in theory.
- They are defined in terms of finite sample bounds and depend on prior assumptions on the problem.
- In practice heuristics are typically adopted.
- In practice many regularization parameter need to be chosen.

Parameter choice is a HUGE unsolved problem.