On Tikhonov Regularization Algorithms

in Learning Theory

Andrea Caponnetto (DISI, Genova University)
Lorenzo Rosasco (DISI)
Ernesto Devito (Modena University)
Michele Piana (DIMA, Genova University)
Alessandro Verri *(DISI)*

Plan

- Notation
- Part I:
  - general form of the solution of Tikhonov algorithms
  - existence and uniqueness
- Part II: a data independent bound on the solution (for the square loss)

Ingredients

1. The **sample space** $Z = X \times Y$, with X a closed subset of $\mathbb{R}^n$ and Y a closed subset of $\mathbb{R}$.

2. The **probability measure** ρ on the sample space Z.

3. The **training set** $D = ((x_1, y_1), \cdots, (x_\ell, y_\ell))$, a sequence ℓ examples drawn i.i.d. according to the probability ρ. Z.

4. **Regression**: the **labels** y belong to $\mathbb{R}$; **Classification** $y = \pm 1$

More ingredients

The **loss function** $V(y, f(x))$ is the price we are willing to pay by using $f(x)$ to predict the correct label y.

The **expected risk**, defined as

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

can be seen as the average error obtained by a solution f of the learning problem.

Given a training set D the **empirical risk** is defined as

$$I^D_{emp}[f] = \frac{1}{\ell} \sum_{i=1}^{\ell} V(y_i, f(x_i))$$

The Learning Problem

The problem of learning is to find, given the training set D, an **estimator** f effectively predicting the label of a **new** point.

This translates in finding a function f such that its expected risk is small with high probability.

Tikhonov Regularization

A possible way to efficiently solve the learning problem is provided by **Regularization Networks** (Girosi and Poggio 92, Evgeniou et al 2000) which amounts to solve the following minimization problem

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

where V is the loss function, $\mathcal{H}$ is the **Hypothesis space**, $\lambda > 0$ is the **regularization parameter** and $(x_i, y_i)_{i=1}^{\ell}$ are the ℓ pairs of examples.

Previous Work: Representer Theorem

If we let $\mathcal{H}$ be RKHS, it can be shown (Wahba 70, Wahba90, Girosi et al 95, Scholkopf et al. 01) that, if a solution exists, it can be written as

$$f_D^\lambda(x) = \sum_{i=1}^{\ell} \alpha_i K(x, x_i)$$

It is also interesting to consider the case in which an offset term b appears in the explicit form of the solution

$$f_D^\lambda(x) = \sum_{i=1}^{\ell} \alpha_i K(x, x_i) + b$$

Tikhonov Functional in the Continuous Setting

We study the following functional

$$\min_{(f,g)\in\mathcal{H}\times\mathcal{B}} \int_{X\times Y} V(y, f(x)+g(x))d\rho(x,y) + \lambda \|f\|^2_{\mathcal{H}}.$$

Minimization takes place in the set $\mathcal{H}\times\mathcal{B}$, where $\mathcal{H}$ and $\mathcal{B}$ are RKHS with kernel K and $K^{\mathcal{B}}$ respectively.

If we consider the empirical measure

$$\rho_S = \frac{1}{\ell}\sum_{i=1}^{\ell} \delta_{(x_i,y_i)}$$

this reduces to the standard Regularization Network framework.

Hypotheses

Loss function

V is a map $V : Y \times \mathbb{R} \to [0, +\infty[$ such that

1. $\forall\, y \in Y$, $V(y, \cdot)$, is a convex function on $\mathbb{R}$, and continuous on $Y \times \mathbb{R}$

2. there are $b \in [0, +\infty[$ and $a : Y \to \mathbb{R}$ such that

$$V(y, w) \leq a(y) + b|w|^2 \qquad \forall w \in \mathbb{R},\ y \in Y$$

$$\int_{X \times Y} |a(y)| d\rho(x, y) \ < \ +\infty,$$

Hypotheses (cont'd)

Kernels

Since we assume X and Y to be just closed sets we have to require the following conditions

$$\int_{X \times Y} K(x,x) d\rho(x,y) < +\infty$$

$$\int_{X \times Y} K^{\mathcal{B}}(x,x) d\rho(x,y) < +\infty.$$

This ensure that $\mathcal{H}$ and $\mathcal{B}$ can be considered as subspaces of $L^2(Z,\rho)$ and is always true if X is compact or the kernel bounded.

A Quantitative Representer Theorem

Theorem 1
Consider the minimization problem

$$\min_{(f,g)\in\mathcal{H}\times\mathcal{B}} \int_{X\times Y} V(y, f(x) + g(x))d\rho(x,y) + \lambda \|f\|_{\mathcal{H}}^2 .$$

A pair $(f^\lambda, g^\lambda) \in \mathcal{H} \times \mathcal{B}$ is a solution **iff** there are $g^\lambda \in \mathcal{B}$ and $f^\lambda \in \mathcal{H}$ such that

$$f^\lambda = -\frac{1}{2\lambda}\int_{X\times Y} \alpha(x,y) K_x d\rho(x,y),$$

with $\alpha \in L^2(Z, \rho)$, satisfying

$$\alpha(x,y) \in (\partial V)(y, f^\lambda(x) + g^\lambda(x)) \quad \rho\text{−a.e.}$$

$$\int_{X\times Y} \alpha(x,y) K_x^{\mathcal{B}} d\rho(x,y) = 0.$$

Dealing with the Bias Term

The set $\mathcal{H} \times \mathcal{B}$ is not a RKHS (the intersection between $\mathcal{H}$ and $\mathcal{B}$ in not necessarily empty). This makes it difficult to extend typical statistical learning analysis to the setting in which a bias term is considered.

The fact that the estimator is $f^\lambda(x) + g^\lambda(x)$ (for regression) or $\mathrm{sgn}\Big(f^\lambda(x) + g^\lambda(x)\Big)$ (for classification) suggests to replace $\mathcal{H} \times \mathcal{B}$ with the sum

$$\mathcal{S} = \mathcal{H} + \mathcal{B} = \{f + g \in \mathcal{C}(X) \,|\, f \in \mathcal{H},\ g \in \mathcal{B}\}.$$

which is RKHS with kernel $K^\mathcal{S}$ given by the sum $K + K^\mathcal{B}$

Offset Function Space and RKHS

Theorem 2
Let Q be the orthogonal projection on the closed subspace of $\mathcal{S}$

$$\mathcal{S}_0 = \{s \in \mathcal{S} \mid \langle s, g \rangle_{\mathcal{S}} = 0 \;\; \forall g \in \mathcal{B}\},$$

We have the following facts.

1. If $(f^\lambda, g^\lambda) \in \mathcal{H} \times \mathcal{B}$ is a solution of the problem

   $$\min_{(f,g) \in \mathcal{H} \times \mathcal{B}} \{I[f+g] + \lambda \|f\|_{\mathcal{H}}^2\},$$

   then $s^\lambda = f^\lambda + g^\lambda \in \mathcal{S}$ is a solution of the problem

   $$\min_{s \in \mathcal{S}} \{I[s] + \lambda \|Qs\|_{\mathcal{S}}^2\}$$

   and $f^\lambda = Qs^\lambda$.

2. If $s^\lambda \in \mathcal{S}$ is a solution of the problem

   $$\min_{s \in \mathcal{S}} \{I[s] + \lambda \|Qs\|_{\mathcal{S}}^2\},$$

   let $f^\lambda = Qs^\lambda$ and $g^\lambda = s^\lambda - Qs^\lambda$, then

   $$I[f^\lambda + g^\lambda] + \lambda \left\|f^\lambda\right\|_{\mathcal{H}}^2 = \inf_{(f,g) \in \mathcal{H} \times \mathcal{B}} \{I[f+g] + \lambda \|f\|_{\mathcal{H}}^2\}.$$

Comments

- Quantitative version of the representer theorem: very general, it holds for both regression and classification without assuming differentiability of the loss function

- The RKHS sum of the two RKHSs, $\mathcal{H}$ and $\mathcal{B}$, is the natural hypothesis space. The minimization of the Tikhonov functional in $\mathcal{H} \times \mathcal{B}$ is equivalent to the minimization of a Tikhonov functional in which the penalty term is a seminorm.

Existence of the Regularized Solution

If $\mathcal{B} = \emptyset$ the existence is easy to prove. If $\mathcal{B} = \mathbb{R}$ (constant offset functions) existence is ensured by requiring some weak assumptions on the loss function

$\diamondsuit$ **regression**
$\lim_{w\to\pm\infty}(\inf_{y\in Y} V(y, w)) = +\infty$

$\diamondsuit$ **classification**
$\lim_{w\to-\infty} V(1, w) = +\infty$ and $\lim_{w\to+\infty} V(-1, w) = +\infty$

and the kernel

$\diamondsuit$ there is $C > 0$ such that $\sqrt{K(x, x)} \leq C$ for all $x \in \operatorname{supp} \nu$

For classification one must also require to have at least one example for each class.

Uniqueness of the Regularized Solution

For strictly convex functions the uniqueness is ensured if the offset space is *small* enough.

An example of convex loss function which is not strictly convex is the hinge loss of SVM for classification. In this case the solution is unique unless a special condition on the number and location of support vectors is met.

Discrete Setting

Since

$$(\partial V)(y,w) = [V'_-(y,w), V'_+(y,w)],$$

we have that the minimizer of

$$\min_{s \in \mathcal{S}} \left(\frac{1}{\ell} \sum_i V(y_i, s(x_i)) + \lambda \|Qs\|_{\mathcal{S}}^2 \right)$$

can be written as

$$s^\lambda = \sum_{i=1}^{\ell} \alpha_i y_i K_{x_i} + b^\lambda$$

where

$$\begin{aligned} \frac{-1}{2\lambda\ell} V'_+(y_i, f^\lambda(x_i) + b^\lambda) \leq \;\; & \alpha_i \;\; \leq \frac{-1}{2\lambda\ell} V'_-(y_i, f^\lambda(x_i) + b^\lambda) \\ \sum_{i=1}^{\ell} \alpha_i \;\; & = \;\; 0 \end{aligned}$$

Hinge loss: SVM Classification

For the SVM algorithm the conditions on $(\alpha_1, \ldots, \alpha_\ell, b^\lambda)$ translate in the following system of algebraic inequalities

$$
\begin{aligned}
0 \leq \alpha_i \leq C \quad & \text{if} \quad y_i \left(\sum_{j=1}^{\ell} \alpha_j y_j K(x_i, x_j) + b^\lambda \right) = 1 \\
\alpha_i = 0 \quad & \text{if} \quad y_i \left(\sum_{j=1}^{\ell} \alpha_j y_j K(x_i, x_j) + b^\lambda \right) > 1 \\
\alpha_i = C \quad & \text{if} \quad y_i \left(\sum_{j=1}^{\ell} \alpha_j y_j K(x_i, x_j) + b^\lambda \right) < 1 \\
\sum_i \alpha_i y_i \quad & = \quad 0
\end{aligned}
$$

usually obtained as the Kuhn-Tucker conditions of a QP optimization problem

An Example: SVM Classification (cont'd)

It is immediate to establish a link between the form of the loss and the solution properties. The box constraints ($0 \leq \alpha_i \leq C$) are due to the fact that $V(yf(x))$ has an asymptote for $yf(x) \to -\infty$, whereas sparsity ($\alpha_i = 0$) follows from $V(yf(x))$ being constant for $yf(x) > 1$.

Comments

- The offset makes life difficult for both existence and uniqueness

- For constant offsets existence and uniqueness are obtained adding some mild conditions. Convexity of the loss is not sufficient for uniqueness (though in practice it is very likely to be)

- The fact that the Kuhn-Tuker conditions can be obtained in the primal formulation may be useful for understanding other support vector methods and proposing new computational methods

Back to the learning problem (discrete setting)

The problem of learning is to find, given the training set D, an **estimator** f effectively predicting the label of a **new** point.

This translates in finding a function f such that its risk is small with high probability.

A bound for Regularized Least Square RLS

From now on we will focus on the following RLS algorithm. The estimator f_D^λ is defined as the unique solution of the minimization problem

$$\min_{f \in \mathcal{H}} \left(\frac{1}{\ell} \sum_{i=1}^{\ell} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \right).$$

We thus restrict our analysis to the square loss.

Generalization and Model Selection

Model Selection: choose a value λ_0 such that $I[f_D^{\lambda_0}]$ is small with high probability

A possible criterion: given a probabilistic bound of the form

$$\operatorname*{Prob}_{D \in Z^\ell}\left(I[f_D^\lambda] \geq E(\lambda, \eta, \ell, D)\right) \leq \eta$$

for a fixed confidence level $1 - \eta$, choose λ_0 according to the following rule

$$\lambda_0(\eta, \ell, D) = \operatorname*{argmin}_{\lambda > 0}\{E(\lambda, \eta, \ell, D)\},$$

Example of bounds

We distinguish between two type of bounds:

Type 1 $E(\lambda, \eta, \ell, D) = I_{emp}^{D}[f_{D}^{\lambda}] + \Phi(\ell, \eta, \lambda)$

where $\Phi(\ell, \eta, \lambda)$ is a stability or complexity term. (Vapnik 1998, Bousquet et al. 2001...)

Type 2 $E(\lambda, \eta, \ell) = S(\lambda, \eta, \ell) + A(\lambda)$

where S is the sample error due to finite sampling and A is the approximation error due to the fact that we are working in a given Hypothesis space (Vapnik 1998, Cucker et al. 2002...).

Risk of data dependency

Data-dependent bounds introduce a dependency on the training set D in the selected model λ_0.

$$D \Longrightarrow \lambda_0(\eta, D) \Longrightarrow f_D^{\lambda_0(\eta,D)}.$$

It could happen that

$$\operatorname*{Prob}_{D \in Z^\ell}\left(I[f_D^{\lambda_0(\eta,D)}] \geq E(\lambda_0(\eta, D), \eta, D)\right) \gg \eta.$$

Concentration inequality (Mc Diarmid, 1989)

- Let D^i be the training set with the i^{th} example replaced by (x_i', y_i'),

- let ξ be a random variable, $\xi : Z^\ell \to \mathbb{R}$,

- assume that there exists constants $c_i\,(i = 1, \cdots, \ell)$ such that

$$\sup_{D \in Z^\ell} \sup_{(x_i', y_i') \in Z} |\, \xi(D) - \xi(D^i) \,| \leq c_i,$$

then Mc Diarmid inequality gives

$$\operatorname*{Prob}_{D \in Z^\ell} (|\, \xi(D) - E_D(\xi) \,| \geq \epsilon) \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^{\ell} c_i^2}}.$$

A new bound on the expected risk

We consider the real random variable $\xi(D) = \sqrt{I[f_D^\lambda] - \inf_{f \in \mathcal{H}} I[f]}$ and proceed through the following steps

1. estimate the stability of $\xi(D)$ under variations of a single data in the training set D,
2. bound the mean value $E_D\left(\xi(D)\right)$,
3. fix a confidence level $1 - \eta$,
4. apply Mc Diarmid inequality to $\xi(D)$.

Stability of RN

The following strong stability result holds

$$\left| \sqrt{I[f_D^\lambda] - \inf_{f \in \mathcal{H}} I[f]} - \sqrt{I[f_{D^i}^\lambda] - \inf_{f \in \mathcal{H}} I[f]} \right| \leq \frac{2\delta\,\kappa^2}{\lambda \ell} \left(1 + \frac{\kappa}{2\sqrt{\lambda}} \right) =: \frac{1}{\ell} A,$$

where

$$\begin{aligned} \kappa &= \sup\{\sqrt{K(x,x)} \,|\, x \in X\}, \\ \delta &= \sup\{|y| \,|\, y \in Y\}. \end{aligned}$$

$O(\ell^{-1})$ dependency is critical for exponential convergence in the concentration inequality.

The mean value of ξ

It holds

$$\left| E_D\left(\sqrt{I[f_D^\lambda] - \inf_{f\in\mathcal{H}} I[f]}\right) - \sqrt{I[f^\lambda] - \inf_{f\in\mathcal{H}} I[f]}\right| \leq \frac{\kappa^2\delta}{\lambda\sqrt{\ell}}\left(1 + \frac{\kappa}{2\sqrt{\lambda}}\right),$$

where κ and δ are defined as above and f^λ is given by

$$f^\lambda = \operatorname*{argmin}_{f\in\mathcal{H}}\{I[f] + \lambda\|f\|^2_{\mathcal{H}}\}.$$

The term $I[f_D^\lambda]$ can be thought of as *approximation error*. It is the minimum expected risk achievable within the ball of radius $\left\|f^\lambda\right\|_{\mathcal{H}}$ in the RKHS.

The result

Given $0 < \eta < 1$ and $\lambda > 0$, with probability at least $1 - \eta$ it holds

$$\sqrt{I[f_D^\lambda] - \inf_{f \in \mathcal{H}} I[f]} \leq \sqrt{I[f^\lambda] - \inf_{f \in \mathcal{H}} I[f]} + S(\lambda, \eta, \ell),$$

where

$$S(\lambda, \eta, \ell) \;=\; \frac{\delta\,\kappa^2}{\lambda\sqrt{\ell}}\left(1 + \frac{\kappa}{2\sqrt{\lambda}}\right)(1 + \sqrt{2}\log\frac{1}{\eta}).$$

The term $S(\lambda, \eta, \ell)$ plays the role of *sample error*. It measures the deviation due to finite sampling, of $I[f_D^\lambda]$ from the approximation error.

Conclusions

1. Data dependent bounds are risky

2. We derived a bound using stability of RLS

3. It can also be shown that the proposed RLS algorithm is consistent, because for every $\epsilon > 0$ it holds

$$\lim_{\ell \to \infty} \text{Prob}\{D \in Z^\ell \,|\, I[f_D^{\lambda_0(\ell)}] > \inf_{f \in \mathcal{H}} I[f] + \epsilon\} = 0.$$

Strong Consistency in Probability

Definition: The one parameter family of estimators $\{f_S^\lambda\}_\lambda$ provided with a model selection rule $\lambda_0(\ell)$ is strongly consistent in probability iff, for every $\epsilon > 0$ it holds

$$\lim_{\ell \to \infty} \mathrm{Prob}\{D \in Z^\ell \,|\, I[f_D^{\lambda_0(\ell)}] > \inf_{f \in \mathcal{H}} I[f] + \epsilon\} = 0.$$

Consistency Results

1. We defined the regularization parameter λ_0 as a function of the number of examples ℓ and the confidence level $1 - \eta$,

$$\lambda_0(\ell, \eta) = \max \operatorname*{argmin}_{\lambda \in [0, +\infty]} E(\lambda, \eta, \ell).$$

2. We now define a model selection rule only depending on ℓ by introducing a power-law dependency of η on ℓ,

$$\lambda_0(\ell) = \lambda_0(\ell, \ell^{-p}), \text{ with } p > 0.$$

3. It can be proved that the sequence $(\lambda_0(\ell))_{\ell=1}^{\infty}$ *is not increasing, tends to zero and provides strong consistency in probability*.

Concluding...