Regularization

Regularization is the classical way to restore well posedness (and ensure generalization). Regularization in general means retricting H, as we have in fact done for ERM. There are two standard approaches in the field of ill-posed problems that ensure for ERM *well-posedness* (and *generalization*) by constraining the hypothesis space $\mathcal{H}$. The direct way – minimize the empirical error subject to f in a ball in an appropriate $\mathcal{H}$ – is called *Ivanov regularization*. The indirect way is *Tikhonov regularization* (which is not strictly ERM).

Ivanov and Tikhonov Regularization

ERM finds the function in $(\mathcal{H})$ which minimizes

$$\frac{1}{n}\sum_{i=1}^{n} V(f(x_i), y_i)$$

which in general – for arbitrary hypothesis space $\mathcal{H}$ – is *ill-posed*.

- Ivanov regularizes by finding the function that minimizes

$$\frac{1}{n}\sum_{i=1}^{n} V(f(x_i), y_i)$$

  while satisfying $\mathcal{R}(f) \leq A$.

- Tikhonov regularization minimizes over the hypothesis space $\mathcal{H}$, for a fixed positive parameter γ, the regularized functional

$$\frac{1}{n}\sum_{i=1}^{n} V(f(x_i), y_i) + \gamma\mathcal{R}(f). \tag{2}$$

$\mathcal{R}(f)$ is the regulirizer, a penalization on f. In this course we will mainly discuss the case $\mathcal{R}(f) = \|f\|_K^2$ where $\|f\|_K^2$ is the norm in the Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$, defined by the kernel K.

Tikhonov Regularization

As we will see in future classes

- Tikhonov regularization ensures well-posedness eg existence, uniqueness and especially *stability* (in a very strong form) of the solution
- Tikhonov regularization ensures generalization
- Tikhonov regularization is closely related to – but different from – Ivanov regularization, eg ERM on a hypothesis space $\mathcal{H}$ which is a ball in a RKHS.

Remarks on Foundations of Learning Theory

Intelligent behavior (at least learning) consists of optimizing under constraints. Constraints are key for solving computational problems; constraints are key for prediction. Constraints may correspond to rather general symmetry properties of the problem (eg time invariance, space invariance, invariance to physical units (pai theorem), universality of numbers and metrics implying normalization, etc.)

- Key questions at the core of learning theory:
  - generalization and predictivity *not* explanation
  - probabilities are unknown, only data are given
  - which constraints are needed to ensure generalization (therefore which hypotheses spaces)?
  - regularization techniques result usually in *computationally "nice"* and *well-posed* optimization problems

Statistical Learning Theory and Bayes

Unlike statistical learning theory the Bayesian approach does not emphasize

- the issue of generalization (following the tradition in statistics of explanatory statistics);
- that probabilities are not known and that only data are known: assuming a specific distribution is a very strong – unconstrained by any Bayesian theory – seat-of-the-pants guess;
- the question of which priors are needed to ensure generalization;
- that the resulting optimization problems are often *computationally intractable* and possibly ill-posed optimization problems (for instance not unique).

Plan

- Part I: Basic Concepts and Notation
- Part II: Foundational Results
- Part III: Algorithms

Hypotheses Space

We are going to look at hypotheses spaces which are reproducing kernel Hilbert spaces.

- RKHS are **Hilbert spaces** of **point-wise defined** functions.
- They can be defined via a **reproducing kernel**, which is a symmetric positive definite function.

$$\sum_{i,j=1}^{n} c_i c_j K(t_i, t_j) \geq 0$$

for any $n \in \mathbb{N}$ and choice of $t_1, ..., t_n \in X$ and $c_1, ..., c_n \in \mathbb{R}$.

- functions in the space are (the completion of) linear combinations

$$f(x) = \sum_{i=1}^{p} K(x, x_i) c_i.$$

- the norm in the space is a natural measure of complexity

$$\sum_{i,j=1}^{p}$$

Hypotheses Space

We are going to look at hypotheses spaces which are reproducing kernel Hilbert spaces.

- RKHS are **Hilbert spaces** of **point-wise defined** functions.
- They can be defined via a **reproducing kernel**, which is a symmetric positive definite function.

$$\sum_{i,j=1}^{n} c_i c_j K(t_i, t_j) \geq 0$$

for any $n \in \mathbb{N}$ and choice of $t_1, ..., t_n \in X$ and $c_1, ..., c_n \in \mathbb{R}$.

- functions in the space are (the completion of) linear combinations

$$f(x) = \sum_{i=1}^{p} K(x, x_i) c_i.$$

- the norm in the space is a natural measure of complexity

$$p$$

Hypotheses Space

We are going to look at hypotheses spaces which are reproducing kernel Hilbert spaces.

- RKHS are **Hilbert spaces** of **point-wise defined** functions.
- They can be defined via a **reproducing kernel**, which is a symmetric positive definite function.

$$\sum_{i,j=1}^{n} c_i c_j K(t_i, t_j) \geq 0$$

for any $n \in \mathbb{N}$ and choice of $t_1, ..., t_n \in X$ and $c_1, ..., c_n \in \mathbb{R}$.

- functions in the space are (the completion of) linear combinations

$$f(x) = \sum_{i=1}^{p} K(x, x_i) c_i.$$

- the norm in the space is a natural measure of complexity

Hypotheses Space

We are going to look at hypotheses spaces which are reproducing kernel Hilbert spaces.

- RKHS are **Hilbert spaces** of **point-wise defined** functions.
- They can be defined via a **reproducing kernel**, which is a symmetric positive definite function.

$$\sum_{i,j=1}^{n} c_i c_j K(t_i, t_j) \geq 0$$

for any $n \in \mathbb{N}$ and choice of $t_1, ..., t_n \in X$ and $c_1, ..., c_n \in \mathbb{R}$.

- functions in the space are (the completion of) linear combinations

$$f(x) = \sum_{i=1}^{p} K(x, x_i) c_i.$$

- the norm in the space is a natural measure of complexity

$$p$$

Examples of pd kernels

Very common examples of symmetric pd kernels are

- **Linear kernel**

$$K(x, x') = x \cdot x'$$

- **Gaussian kernel**

$$K(x, x') = \mathrm{e}^{-\frac{\|x-x'\|^2}{\sigma^2}}, \qquad \sigma > 0$$

- **Polynomial kernel**

$$K(x, x') = (x \cdot x' + 1)^d, \qquad d \in \mathbb{N}$$

For specific applications, designing an effective kernel is a challenging problem.

Kernel and Features

Often times kernels, are defined through a dictionary of features

$$\mathcal{D} = \{\phi_j,\ i = 1, \ldots, p \mid \phi_j : X \to \mathbb{R},\ \ \forall j\}$$

setting

$$K(x, x') = \sum_{i=1}^{p} \phi_j(x)\phi_j(x').$$

Ivanov regularization

We can regularize by explicitly restricting the hypotheses space $\mathcal{H}$ — for example to a ball of radius R.

Ivanov regularization

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

subject to

$$\|f\|_{\mathcal{H}}^2 \leq R.$$

The above algorithm corresponds to a constrained optimization problem.

Tikhonov regularization

Regularization can also be done implicitly via penalization

Tikhonov regularizarion

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

λ is the regularization parameter trading-off between the two terms.

The above algorithm can be seen as the Lagrangian formulation of a constrained optimization problem.

The Representer Theorem

An important result

The minimizer over the RKHS $\mathcal{H}$, f_S, of the regularized empirical functional

$$I_S[f] + \lambda \|f\|_{\mathcal{H}}^2,$$

can be represented by the expression

$$f_n(x) = \sum_{i=1}^{n} c_i K(x_i, x),$$

for some $(c_1, \ldots, c_n) \in \mathbb{R}$.

Hence, minimizing over the (possibly infinite dimensional) Hilbert space, *boils down to minimizing over* $\mathbb{R}^n$.