Regularization Theory and Machine Learning

PhD Course: Introduction to Inverse Problem

Salvatore Frandina
salvatore.frandina@gmail.com

Department of Information Engineering, Siena, Italy

Siena, August 19, 2012

Summary

An overview of the Regularization Theory

- General Regularization Theory
- Tikhonov's Regularization Theory
- Regularization Theory and Machine Learning

Direct and Inverse Problems

- **Direct problem**: given a know "model" x in a model space X and an operator K, evaluate $K(x)$
- **Inverse problem**: given y and an operator K, solve the equation $K(x) = y$ for determine the "model "x
- **Note**:
  * The formulation of an inverse problem requires the definition of the operator K with its domain and range
  * The formulation as an operator equation allows to distinguish among finite, semi-finite, and infinite-dimensional, linear and nonlinear problems
  * The evaluation of $K(x)$ requires solving a boundary value problem for a differential equation or evaluating an integral equation

Well-Posedness I

- A mathematical model for a physical **problem is well-posed**, in the sense of Hadamard, if it satisfies the following properties:
  1. there exists a solution of the problem (existence)
  2. there is at most one solution of the problem (uniqueness)
  3. the solution depends continuously on the data (stability)
- The direct problem is well-posed while the inverse problem is ill-posed
- **Note**:
  * The existence of a solution can be enforced by enlarging the solution space
  * If a problem has more than one solution, then information about the model is missing
  * If the solution of a problem does not depend continuously on the data, then in general the computed solution has nothing to do with the true solution

Well-Posedness II

Definition 1

Let X and Y be normed spaces, $K : X \to Y$ a (linear or nonlinear) mapping. The equation $Kx = y$ is called well-posed if the following holds:

1. Existence: for every $y \in Y$ there is (at least one) $x \in X$ such that $Kx = y$
2. Uniqueness: for every $y \in Y$ there is at most one $x \in X$ with $Kx = y$
3. Stability: the solution x depends continuously on y; that is, for every sequence $(x_n) \subset X$ with $Kx_n \to Kx$ $(n \to \infty)$, it follows that $x_n \to x$ $(n \to \infty)$

- every equations for which (at least) one of these properties does not hold are called ill-posed

Well-Posedness III

- **Existence** and **Uniqueness** depend only on the algebraic nature of the spaces and the operator i.e. if the operator is onto (surjective) or one-to-one (injective)
- **Stability** depends also on the topologies of the spaces i.e. if the inverse operator $K^{-1} : Y \to X$ is continuous
- All these requirements are not independent of each other. A theorem says that the inverse operator K^{-1} is automatically continuous if K is linear and continuous and X and Y are Banach spaces

Compact Operator

- Integral operators are compact operators in many natural topologies under very weak conditions on the kernels

Theorem 1

Let X, Y be normed spaces and $K : X \to Y$ be a linear compact operator with kernel $\mathcal{N}(K) := \{x \in X : Kx = 0\}$. Let the dimension of the factor space $X/\mathcal{N}(K)$ be infinite. Then there exists a sequence $(x_n) \subset X$ such that $Kx_n \to 0$ but (x_n) does not converge. We can even choose (x_n) such that $\|x_n\| \to \infty$. In particular, if K is one-to-one, the inverse $K^{-1} : Y \supset \mathcal{R}(K) \to X$ is unbounded. Here, $\mathcal{R}(K) := \{Kx \in Y : x \in X\}$ denotes the range of K

- Linear equations of the form $Kx = y$ with compact operators K are always ill-posed

General Concept of Regularization I

- Many inverse problems can be formulated as operator equations of the form

$$Kx = y$$

where K is a linear compact operator between Hilbert spaces X and Y over the field $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$

- A successful reconstruction strategy requires additional a priori information about the solution
- The regularization strategies are optimal if they have the same asymptotic order as the worst-case error

General Concept of Regularization II

- The Tikhonov's method and the Landweber's Iteration are two of the most important regularization strategies
- The regularization parameter $\alpha = \alpha(\delta)$ is chosen a priori i.e. before to start to compute the regularized solution
- The optimal regularization parameter α depends on bounds of the exact solution and it is not known in advance

Assumptions of the problem I

- The compact operator K is one-to-one
  - this is not a restriction since that the domain X can be replaced by orthogonal complement of the kernel of K i.e. $N(K)^{\perp} = \{x \in X : \langle x, y \rangle = 0, \ \forall y \in N(K)\}$
- Exists a solution $x \in X$ of the unperturbed equation $Kx = y$ i.e. $y \in R(K) \subset Y$
- The injectivity of K implies that this solution is unique
- The right-hand side $y \in Y$ is never known exactly but only with an error of $\delta > 0$. The assumption is that the $\delta > 0$ and $y^{\delta} \in Y$ are known with $\|y - y^{\delta}\| \leq \delta$

Assumptions of the problem II

- The aim is to "solve" the perturbed equation $Kx^\delta = y^\delta$
- In general, the problem is not solvable because is not known if the measured data y^δ are in the range $\mathcal{R}(K)$ of K
- An approximation $x^\delta \in X$ to the exact solution x must have the same asymptotic error of the worst-case error
- The approximate solution x^δ should depend continuously on the data y^δ

Regularization Strategy I

- The purpose is to construct a suitable bounded approximation $R : Y \to X$ of the (unbounded) inverse operator $K^{-1} : R(K) \to X$

Definition 2

A regularization strategy is a family of linear and bounded operators

$$R_\alpha : Y \to X, \ \alpha > 0,$$

such that for all $x \in X$

$$\lim_{\alpha \to 0} R_\alpha K x = x$$

i.e. the operators $R_\alpha K$ converge pointwise to the identity

Regularization Strategy II

- The following theorem derives the definition 2 and from the compactness of K

Theorem 2

Let R_α be a regularization strategy for a compact operator $K : X \to Y$ where $dim\ X = \infty$. Then

* the operators R_α are not uniformly bounded i.e. there exists a sequence (α_j) with $\|R_{\alpha_j}\| \to \infty$ for $j \to \infty$
* the sequence $(R_\alpha Kx)$ does not converge uniformly on bounded subsets of X i.e. there is no convergence $R_\alpha K$ to the identity I in the operator norm

- N.B. the definition is based on unperturbed data i.e. the regularizer $R_\alpha y$ converges to x for the exact right-hand side $y = Kx$

Regularization Strategy III

- An approximation of the solution x of $Kx = y$ is defined as

$$x^{\alpha,\delta} = R_\alpha y^\delta$$

  where $y \in K(X)$ is the exact right-hand side and $y^\delta \in Y$ is the measured data with $\|y - y^\delta\| \leq \delta$
- The error can be split into two parts using triangle inequality

$$\|x^{\alpha,\delta} - x\| \leq \|R_\alpha y^\delta - R_\alpha y\| + \|R_\alpha y - x\| \leq \|R_\alpha\| \|y^\delta - y\| + \|R_\alpha K x - x\|$$

  and finally

$$\|x^{\alpha,\delta} - x\| \leq \underbrace{\delta}_{\text{error in the data}} \overbrace{\|R_\alpha\|}^{\text{condition number}} + \underbrace{\|R_\alpha K x - x\|}_{\text{approximation error}} \qquad (1)$$

Total Error Trend

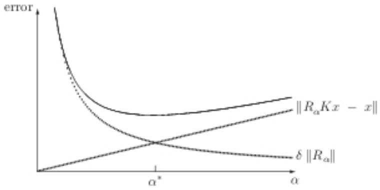

Figure: Trend of the total error

- The first term is the error in the data multiplied by the "condition number" $\|R_\alpha\|$ of the regularized problem. By theorem 2, this term tends to infinity as α tends to zero
- The second term is the approximation error $\|(R_\alpha - K^{-1})y\|$ at the exact right-hand side $y = Kx$. By the definition of a regularization strategy, this term tends to zero with α
- The aim is to choose $\alpha = \alpha(\delta)$ dependent on δ in order to minimize the total error $\delta\|R_\alpha\| + \|R_\alpha Kx - x\|$

Admissible Regularization Strategy

Definition 3

A regularization strategy $\alpha = \alpha(\delta)$ is called admissible if, for every $x \in X$, $\alpha(\delta) \to 0$ and

$$\sup\{\|R_{\alpha(\delta)}y^{\delta} - x\| : \|Kx - y^{\delta}\| \leq \delta\} \to 0, \ \delta \to 0$$

- A method to construct classes of admissible regularization strategies is given by **filtering singular systems**
- Let $K : X \to Y$ be a linear compact operator, and let (μ_j, x_j, y_j) be a singular system for K. The solution x of $Kx = y$ is given by Picard's theorem [1] as

$$x = \sum_{j=1}^{\infty} \frac{1}{\mu_j}(y, y_j)x_j$$

provided the series converges i.e. $y \in R(K)$

Regularizing Filter I

Theorem 3

Let $K : X \to Y$ be a compact with singular system (μ_j, x_j, y_j) and $q : (0, \infty) \times (0, \|K\|] \to \mathbb{R}$ be a function with the following properties:

- (1) for every $\alpha > 0$ and for all $0 < \mu \leq \|K\|$
  - $|q(\alpha, \mu)| \leq 1$
  - exists $c(\alpha)$ such that $|q(\alpha, \mu)| \leq c(\alpha)\mu$
- (2) for every $0 < \mu \leq \|K\|$, $\lim_{\alpha \to 0} q(\alpha, \mu) = 1$

Then the operator $R_\alpha : Y \to X$, $\alpha > 0$, defined by

$$R_\alpha y = \sum_{j=1}^{\infty} \frac{q(\alpha, \mu_j)}{\mu_j}(y, y_j)x_j, \; y \in Y,$$

is a regularization strategy with $\|R_\alpha\| \leq c(\alpha)$. A choice $\alpha = \alpha(\delta)$ is admissible if $\alpha(\delta) \to 0$ and $\delta c(\alpha(\delta)) \to 0$ as $\delta \to 0$. The function q is a regularizing filter for K

- the theorem implies that $R_\alpha y$ converges to the solution x

Regularizing Filter II

- A optimal strategy in the sense of worst-case error can be achieved by enforcing the previous theorem 3

Theorem 4

Let assumption (1) of the previous theorem 3 hold. Then (2) can be replaced by the stronger assumption:

- (2a) exists $c_1 > 0$ such that for all $\alpha > 0$ and $0 < \mu \leq \|K\|$ with $|q(\alpha, \mu) - 1| \leq c_1 \frac{\sqrt{\alpha}}{\mu}$
  - If, furthermore, $x \in K^*(Y)$, then
  $$\|R_\alpha Kx - x\| \leq c_1 \sqrt{\alpha} \|z\|,$$
  where $x = K^* z$
- (2b) exists $c_2 > 0$ such that for all $\alpha > 0$ and $0 < \mu \leq \|K\|$ with $|q(\alpha, \mu) - 1| \leq c_2 \frac{\alpha}{\mu^2}$
  - If, furthermore, $x \in K^* K(X)$
  $$\|R_\alpha Kx - x\| \leq c_2 \alpha \|z\|,$$
  where $x = K^* K z$

Regularizing Filter III

Theorem 5

The following three functions q satisfy the assumptions (1), (2), and (2a-b) of the previous theorems 3 and 4:

- (a) $q(\alpha,\mu)=\mu^2/(\alpha+\mu^2)$. This satisfies (1) with $c(\alpha)=1/(2\sqrt{\alpha})$ while assumptions (2a) and (2b) hold with $c_1=1/2$ and $c_2=1$, respectively
- (b) $q(\alpha,\mu)=1-(1-a\mu^2)^{1/\alpha}$ for some $0<a<1/\|K\|^2$. In this case (1) holds with $c(\alpha)=\sqrt{a/\alpha}$ while (2a) and (2b) are satisfied with $c_1=1/\sqrt{2a}$ and $c_2=1/a$, respectively
- (c) Let q be defined by

$$q(\alpha,\mu)=\begin{cases}1, & \mu^2\geq\alpha,\\ 0, & \mu^2<\alpha\end{cases}$$

  In this case (1) holds with $c(\alpha)=1/\sqrt{\alpha}$ while (2a) and (2b) are satisfied with $c1=c2=1$

All of the functions q defined in (a), (b), and (c) are regularizing filters that lead to optimal regularization strategies

Spectral Cutoff

- For the first two choices of q exists a characterization that avoids knowledge of the singular system
- The choice (c) of q is called the **spectral cutoff**. The spectral cutoff solution $x^{\alpha,\delta} \in X$ is therefore defined by

$$x^{\alpha,\delta} = \sum_{\mu_j^2 \geq \alpha} \frac{1}{\mu_j}(y^\delta, y_j)x_j$$

- From the fundamental estimate Definition 1 and with the previous theorem, the following result for the cutoff solution hold

Spectral Cutoff I

Theorem 6

Let $y^\delta \in Y$ be such that $\|y^\delta - y\| \le \delta$, where $y = Kx$ denotes the exact right-hand side

- (a) let $K : X \to Y$ be a compact and injective operator with singular system (μ_j, x_j, y_j). The operators

$$R_\alpha y = \sum_{\mu_j^2 \ge \alpha} \frac{1}{\mu_j}(y, y_j)x_j, \quad y \in Y,$$

  define a regularization strategy with $\|R_\alpha\| \le 1/\sqrt{\alpha}$. This strategy is admissible if $\alpha(\delta) \to 0$ $(\delta \to 0)$ and $\delta^2/\alpha(\delta) \to 0$ $(\delta \to 0)$

- (b) let $x = K^*z \in K^*(Y)$ with $\|z\| \le E$ and $c > 0$. For the choice $\alpha(\delta) = c\delta/E$, we have the estimate

$$\|x^{\alpha(\delta),\delta} - x\| \le \left(\frac{1}{\sqrt{c}} + \sqrt{c}\right)\sqrt{\delta E}$$

- (c) let $x = K^*Kz \in K^*K(X)$ with $\|z\| \le E$ and $c > 0$. The choice $\alpha(\delta) = c(\delta/E)^{2/3}$ leads to the estimate

$$\|x^{\alpha(\delta),\delta} - x\| \le \left(\frac{1}{\sqrt{c}} + c\right)\delta^{2/3}E^{1/3}$$

Spectral Cutoff II

- The spectral cutoff is optimal for the information $\|(K^*)^{-1}x\| \leq E$ or $\|(K^*K)^{-1}x\| \leq E$, respectively (if K^* is one-to-one)

Introduction to Tikhonov's Regularization I

- In general for a concrete integral operators it is recommended to avoid the computation of a singular system
- Given an overdetermined finite linear system of the form $Kx = y$, a common method to solve it is to determine the best fit that minimizes the defect $\|Kx - y\|$ with respect to $x \in X$ for some norm in Y

Introduction to Tikhonov's Regularization II

- If X is infinite-dimensional and K is compact, this minimization problem is also ill-posed by the following lemma

Lemma 1

Let X and Y be Hilbert spaces, $K : X \to Y$ be linear and bounded, and $y \in Y$. There exists $\hat{x} \in X$ with $\|K\hat{x} - y\| \leq \|Kx - y\|$ for all $x \in X$ if and only if $\hat{x} \in X$ solves the normal equation $K^*K\hat{x} = K^*y$. Here, $K^* : Y \to X$ denotes the adjoint of K

Tikhonov functional I

- The aim is to:
  * penalize the defect in term of the optimization theory
  * replace the equation of the first kind $K^*K\hat{x} = K^*y$ with an equation of the second kind in term of language of integral equation theory
- Both viewpoints lead to the following minimization problem: given the linear, bounded operator $K : X \to Y$ and $y \in Y$, determine $x^\alpha \in X$ that minimizes the **Tikhonov functional**

$$J_\alpha(x) = \|Kx - y\|^2 + \alpha\|x\|^2, \quad \forall x \in X$$

Tikhonov functional II

Theorem 7

Let $K : X \to Y$ be a linear and bounded operator between Hilbert spaces and $\alpha > 0$. Then the Tikhonov functional J_α has a unique minimum $x^\alpha \in X$. This minimum x^α is the unique solution of the normal equation

$$\alpha x^\alpha + K^* K x^\alpha = K^* y$$

- The solution x^α can be written in the form $x^\alpha = R_\alpha y$ with

$$R_\alpha = (\alpha I + K^* K)^{-1} K^* : Y \to X \tag{2}$$

Tikhonov functional III

- Choosing a singular system (μ_j, x_j, y_j) for the compact operator K, $R_\alpha y$ has the representation

$$R_\alpha y = \sum_{n=0}^{\infty} \frac{\mu_j}{\alpha + \mu_j^2}(y, y_j)x_j = \sum_{n=0}^{\infty} \frac{q(\alpha, \mu_j)}{\mu_j}(y, y_j)x_j, \quad y \in Y$$

with $q(\alpha, \mu) = \mu^2/(\alpha + \mu^2)$

- **Note**: this function q is exactly the filter function (a) of the theorem 5

Tikhonov regularization method I

Theorem 8

Let $K : X \to Y$ be a linear, compact operator and $\alpha > 0$.

- (a) the operator $\alpha I + K^*K$ is boundedly invertible. The operators $R_\alpha : Y \to X$ from 2 form a regularization strategy with $\|R_\alpha\| \leq 1/(2\sqrt{\alpha})$. It is called **the Tikhonov regularization method**. $R_\alpha y^\delta$ is determined as the unique solution $x^{\alpha,\delta} \in X$ of the equation of the second kind

$$\alpha x^{\alpha,\delta} + K^*Kx^{\alpha,\delta} = K^*y^\delta$$

  Every choice $\alpha(\delta) \to 0$ $(\delta \to 0)$ with $\delta^2/\alpha(\delta) \to 0$ $(\delta \to 0)$ is admissible
- (b) let $x = K^*z \in K^*(Y)$ with $\|z\| \leq E$. Choosing $\alpha(\delta) = c\delta/E$ for some $c > 0$. Then the following estimate holds:

$$\|x^{\alpha(\delta),\delta} - x\| \leq \frac{1}{2}\left(\frac{1}{\sqrt{c}} + \sqrt{c}\right)\sqrt{\delta E}$$

- (c) let $x = K^*Kz \in K^*K(X)$ with $\|z\| \leq E$. The choice $\alpha(\delta) = c(\delta/E)^{2/3}$ for some $c > 0$ leads to the error estimate

$$\|x^{\alpha(\delta),\delta} - x\| \leq \left(\frac{1}{2\sqrt{c}} + c\right)\delta^{2/3}E^{1/3}$$

The Tikhonov's regularization method is optimal for the information $(K^*)^{-1}x \leq E$ or $(K^*K)^{-1}x \leq E$, respectively (provided K^* is one-to-one)

Tikhonov regularization method II

- The eigenvalues of K tend to zero, and the eigenvalues of $\alpha I + K^* K$ are bounded away from zero by $\alpha > 0$ [1]
- From the previous theorem 8, the $\alpha(\delta)$ has to be chosen in such a way that it converges to zero as δ tends to zero but not as fast as δ^2
- From parts (b) and (c), the smoother solution x is the slower α has to tend to zero
- N.B. the convergence can be arbitrarily slow if no a priori assumption about the solution x is available [1]

Tikhonov regularization method III

Theorem 9

Let $K : X \to Y$ be linear, compact, and one-to-one such that the range $\mathcal{R}(K)$ is infinite-dimensional. Furthermore, let $x \in X$, and assume that there exists a continuous function $\alpha : [0, \infty) \to [0, \infty)$ with $\alpha(0) = 0$ such that

$$\lim_{\delta \to 0} \|x^{\alpha(\delta),\delta} - x\| \delta^{-2/3} = 0$$

for every $y^\delta \in Y$ with $\|y^\delta - Kx\| \leq \delta$, where $x^{\alpha(\delta),\delta} \in X$ solves $\alpha x^{\alpha,\delta} + K^* K x^{\alpha,\delta} = K^* y^\delta$. Then $x = 0$

Consideration

- Tikhonov's regularization method is not optimal for stronger "smoothness" assumptions on the solution x i.e. under the assumption $x \in (K^*K)^r(X)$ for some $r \in \mathbb{N}$, $r \geq 2$
  * this is in contrast to Landweber's method or the conjugate gradient method
- The choice of α in theorem 8 is made a priori; that is, before starting the computation of x_α by solving the least squares problem
- **Note**: The regularization is stronger related to the differential operators or norms used [1]
  * one example is the interpretation of regularization by smoothing norms in terms of reproducing kernel Hilbert spaces

Landweber's Equation

- Landweber suggested rewriting the equation $Kx = y$ in the form $x = (I - aK^*K)x + aK^*y$ for some $a > 0$ and iterating this equation:
  * $x^0 = 0$
  * $x^m = (I - aK^*K)x^{m-1} + aK^*y$ for $m = 1, 2, \dots$
- This iterative scheme can be viewed as the steepest descent algorithm applied to the quadratic functional $x \to \|Kx - y\|^2$

Landweber's Iteration I

Lemma 2

Consider the sequence $(x^m = (I - aK^*K)x^{m-1} + aK^*y)$ and define the functional $\psi : X \to \mathbb{R}$ by $\psi(x) = \frac{1}{2}\|Kx - y\|^2$, $x \in X$. Then ψ is Fréchet differentiable in every $z \in X$ and

$$\psi^{'}(z)x = Re(Kz - y, Kx) = Re(K^*(Kz - y), x), \;\; x \in X$$

Therefore, $x^m = x^{m-1} - aK^*(Kx^{m-1} - y)$ is **the steepest descent step** with stepsize a

Landweber's Iteration II

- x^m can be now expressed by the explicit form $x^m = R_m y$, where the operator $R_m : Y \to X$ is defined by

$$R_m = a \sum_{k=0}^{m-1} (I - aK^*K)^k K^* \quad \textit{for} \;\; m = 1, 2, \ldots$$

- Choosing a singular system (μ_j, x_j, y_j) for the compact operator K, then $R_m y$ has the representation

$$R_m y = \sum_{j=1}^{\infty} \frac{q(m, \mu_j)}{\mu_j} (y, y_j) x_j, \quad y \in Y$$

with $q(m, \mu) = 1 - (1 - a\mu^2)^m$

- This filter function q is the same of the theorem 5 where $\alpha = 1/m$

Landweber's Iteration III

Theorem 10

- (a) let $K : X \to Y$ be a compact operator and let $0 < a < 1/\|K\|^2$. The linear and bounded operators $R_m : Y \to X$ define a regularization strategy with discrete regularization parameter $\alpha = 1/m$, $m \in N$, and $\|R_m\| \leq \sqrt{am}$. The sequence $x^{m,\delta} = R_m y^\delta$ is computed by the iteration
  - $x^{0,\delta} = 0$
  - $x^{m,\delta} = (I - aK^*K)x^{m-1,\delta} + aK^*y^\delta$

  for $m = 1, 2, \ldots$. Every strategy $m(\delta) \to \infty$ $(\delta \to 0)$ with $\delta^2 m(\delta) \to 0$ $(\delta \to 0)$ is admissible
- (b) let $x = K^*z \in K^*(Y)$ with $\|z\| \leq E$ and $0 < c_1 < c_2$. For every choice $m(\delta)$ with $c_1 \frac{E}{\delta} \leq m(\delta) \leq c_2 \frac{E}{\delta}$, the following estimate holds:

$$\|x^{m(\delta),\delta} - x\| \leq c_3 \sqrt{\delta E}$$

  for some c_3 depending on c_1, c_2, and a.
- (c) let $x = K^*Kz \in K^*K(X)$ with $\|z\| \leq E$ and $0 < c_1 < c_2$. For every choice $m(\delta)$ with $c_1(E/\delta)^{2/3} \leq m(\delta) \leq c_2(E/\delta)^{2/3}$, we have

$$\|x^{m(\delta),\delta} - x\| \leq c_3 E^{1/3} \delta^{2/3}$$

Consideration

- From the previous slide, the Landweber's iteration is optimal for the information $\|(K^*)^{-1}\|x \leq E$ or $\|(K^*K)^{-1}x\| \leq E$, respectively
- The choice $x^0 = 0$ simplifies the analysis. In general, the explicit iteration x^m is given by

$$x^m = a \sum_{k=0}^{m-1} (I - aK^*K)^k K^* y + (I - aK^*K)^m x^0, \quad m = 1, 2, \ldots$$

- R_m **is affine linear** i.e. of the form $R_m y = z^m + S_m y$, $y \in Y$, for some $z^m \in X$ and some linear operator $S_m : Y \to X$
- **Note**:
  * high precision (ignoring the presence of errors) requires a large number m of iterations
  * stability forces us to keep m small enough
  * from the theorem, it also holds the following general rule:
    $x \in (K^*K)^r(X), r \in \mathbb{N}$
    $\|x^{m(\delta),\delta} - x\| \leq c\, E^{\frac{1}{2r+1}}\, \delta^{\frac{2r}{2r+1}}$

Regularization Theory and Machine Learning I

- Most of the inverse problems in science and engineering areas are ill-posed
  * computational vision
  * system identification
  * nonlinear dynamic reconstruction
  * density estimation
- Given the available input data, the solution to the problem is nonunique (one-to-many) or unstable
- The classical regularization techniques, due to Tikhonov, are able to making the solution well-posed
  * model selection
  * complexity control

Regularization Theory and Machine Learning II

- Regularization theory was introduced to the machine learning community in the 1990s
- A regularization algorithm for learning is equivalent to a multilayer network with a kernel in the form of a *radial basis function* (RBF), an RBF network [2]
- The solution of the classical Tikhonov regularization problem can be derived from the regularized functional defined by a:
  * linear differential operator in the spatial domain
  * linear integral operator in Fourier domain
- **Note**: the regularized solution was originally derived by a differential linear operator and its Green's function

Introduction II

- One way to solve ill-posed problems is to make the problems well-posed by incorporating prior knowledge into the solutions
- The forms of prior knowledge vary and are problem dependent
  - the most popular and important prior knowledge is the smoothness prior
- A possible choice of the smoothness of the functional is to put the functional into the *reproducing kernel Hilbert space* (RKHS)

Machine Learning Framework

- Given a set of observation data (learning examples) $\{(\mathbf{x}_i, y_i) \in \mathbb{R}^N \times \mathbb{R}\} \subset X \times Y$, the learning machine f has to find the solution to the inverse problem
- f has to approximate a real function in the hypothesis satisfying the constraints $f(\mathbf{x}_i) = y(\mathbf{x}_i) \equiv y_i$
  * where $y(x)$ is supposed to be a deterministic function in the target space
- The learning problem can be viewed as a multivariate functional approximation problem
- **Note**: this problem is ill-posed since the approximants satisfying the constraints are not unique

Loss Function

- Statistically, the approximation accuracy is measured by the expectation of the approximation error. The expected risk functional

$$\mathcal{R} = \int_{X \times Y} L(\mathbf{x}, y) p(\mathbf{x}, y) d\mathbf{x} dy$$

where $L(\mathbf{x}, y)$ represents the loss functional
- A common loss function is the mean squared error defined by L_2 norm and the expected risk functional is

$$\mathcal{R} = \int_{X \times Y} [y - f(\mathbf{x}, y)]^2 p(\mathbf{x}, y) d\mathbf{x} dy$$

- In practice
  * the joint probability $p(\mathbf{x}, y)$ is unknown
  * an estimate of $\mathcal{R}$ is based on finite l observations and so the expected risk functional becomes an **empirical risk functional**

$$\mathcal{R}_{emp} = \sum_{i=1}^{l} [y_i - f(\mathbf{x}_i)]^2$$

  * which introduces an estimate $\hat{y}(\mathbf{x})$ ($\hat{y} = y - \epsilon = f(\mathbf{x})$)

Tikhonov Regularization

- The expected risk can be decomposed into two part

$$\mathcal{R}[f] = \underbrace{\mathcal{R}_{emp}[f]}_{\text{the empirical risk functional}} + \underbrace{\mathcal{R}_{reg}[f]}_{\text{the regularizer risk functional}}$$

$$= \frac{1}{2}\sum_{i=1}^{l}[y_i - f(\mathbf{x}_i)]^2 + \frac{1}{2}\lambda\|\mathbf{D}f\|^2$$

  * $\|\cdot\|$ is the norm operator
  * λ **is a regularization parameter** that controls the trade-off between the good fit of the data and the smoothness of the solution
  * **D is a linear differential operator** defined as the Fréchet differential of Tikhonov functional

- **Note**: the smoothness prior implicated in **D** makes the solution stable and insensitive to noise

The Fréchet differential and Riesz Theorem

Definition 4

A function $f : X \to Y$ is Fréchet differentiable at a point x, if for every $h \in X$,

$$\lim_{\epsilon \to 0} \frac{f(x + \epsilon h) - f(x)}{\epsilon} = df(x, h)$$

exists, and defines a linear bounded transformation (in h) mapping X into Y. $df(x, h) = F(x)h$ is the Fréchet differential; $F(x)$ is the Fréchet derivative.

Theorem 11

Let H be a Hilbert space, and let H^* be its dual space, consisting of all continuous linear functionals from H into the field $\mathbb{R}$ or $\mathbb{C}$. If x is an element of H, then the function φ_x defined by

$$\varphi_x(y) = \langle y, x \rangle \quad \forall y \in H$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product of the Hilbert space, is an element of H^*. The theorem states that every element of H^* can be written uniquely in this form

The Reproducing Kernel Hilbert Space I

Definition 5

Let X be an arbitrary set and H a Hilbert space of complex-valued functions on X. H is a reproducing kernel Hilbert space if every linear functional of the form $L_x : f \to f(x)$ ($L_x : H \to \mathbb{C}$) is continuous for any $x \in X$. By the Riesz theorem, this implies that for every $x \in X$ there exists a unique element K_x of H with the property that:

$$f(x) = \langle f, K_x \rangle \quad \forall f \in H.$$

The function K_x is called the point-evaluation functional at the point x. The function $K : X \times X \to \mathbb{C}$ defined by $K(x, y) = \overline{K_x(y)}$ is called the reproducing kernel for the Hilbert space H and it is determined entirely by H

The Reproducing Kernel Hilbert Space II

- **Note**:
  * H is a space of functions, then the element K_x is itself a function
  * the Riesz theorem guarantees that, for every x in X, the element K_x is unique
- **The reproducing property**
  * $K(x, y) = \overline{K_x(y)} = \langle K_y, K_x \rangle$
  * $K(x, x) = \langle K_x, K_x \rangle \geq 0, \quad \forall x \in X$
  * $K_x = 0 \quad \text{if and only if} \quad f(x) = 0 \quad \forall f \in H$

Fréchet Differential in Spatial Domain

- The Fréchet Differential of $\mathcal{R}$ is $d\mathcal{R}(f,h) = \frac{d}{d\beta}\mathcal{R}(f+\beta h)\Big|_{\beta=0}$ where $h(\mathbf{x})$ is a constant function of $\mathbf{x}$

$$\mathcal{R}_{emp}(f,h) = \frac{d}{d\beta}\mathcal{R}_{emp}(f+\beta h)\Big|_{\beta=0}$$
$$= -\sum_{i=1}^{l}[y_i - f(\mathbf{x}_i)]h(\mathbf{x}_i) = -\left\langle h, \sum_{i=1}^{l}[y_i - f(\mathbf{x}_i)]\delta(\mathbf{x}-\mathbf{x}_i)\right\rangle$$

$$\mathcal{R}_{reg}(f,h) = \frac{d}{d\beta}\mathcal{R}_{reg}(f+\beta h)\Big|_{\beta=0}$$
$$= \int \mathbf{D}[f+\beta h]\mathbf{D}h \, d\mathbf{x}\Big|_{\beta=0} = \int \mathbf{D}f\mathbf{D}h \, d\mathbf{x} = \langle \mathbf{D}f, \mathbf{D}h\rangle$$

* $\delta(\cdot)$ is the Dirac delta function
* $\langle\cdot,\cdot\rangle$ is the inner product in Hilbert space H

Integral Equation I

- Given a bounded linear operator $A : X \to Y$, its adjoint operator is defined as $A^* : Y \to X$ such that $\langle Ax, y \rangle = \langle x, A^* y \rangle$ $(x \in X, y \in Y)$ and $\|A\| = \|A^*\|$
- An operator is a self-adjoint operator if it is equal to its adjoint operator $\langle Ax, x' \rangle = \langle x, A^* x' \rangle$ $(x, x' \in X)$

Definition 6

Given a positive-definite kernel function K

* an integral operator $\mathbf{T}$ is defined by $\mathbf{T}f = f(s) = \int K(s,x) f(x)\, dx$
* $\mathbf{T}^*$ is the adjoint operator defined by $\mathbf{T}^* f = f(x) = \int K(x,s) f(s)\, ds$
* $\mathbf{T}$ is self-adjoint if and only if $K(s,x) = \overline{K(x,s)}$ for all s and x

- The integral equation is a Fredholm integral equation of the first kind

Integral Equation II

- **Note**:
- An integral operator with a symmetric kernel $K(s, x)$ is self-adjoint
- $K = \mathbf{T}^*\mathbf{T}f = g(s) = \int K'(s, x)f(x)dx$ with $K'(s, x) = \int K(s, t)\overline{K(x, t)}dt$ K is a integral self-adjoint operator
- $\mathbf{L} = \mathbf{D}^*\mathbf{D}$ is a differential operator

Green's Function I

- Given a positive integral operator **T** is possible to find a (pseudo-)differential operator **D** as its inverse
- The operator **D** corresponds to the inner product of the RKHS with a reproducing kernel K associated to **T**
  * the kernel K is called Green's function of the differential operator **D**

Definition 7

Given a linear differential operator **L**, the function $G(\mathbf{x}, \boldsymbol{\xi})$ is the Green's function for **L** if it has the following properties:

* for a fixed $\boldsymbol{\xi}$, $G(\mathbf{x}, \boldsymbol{\xi})$ is a function of $\mathbf{x}$ and satisfies the given boundary conditions
* except at the point $\mathbf{x} = \boldsymbol{\xi}$, the derivatives of $G(\mathbf{x}, \boldsymbol{\xi})$ with respect to $\mathbf{x}$ are all continuous; the number of derivatives is determined by the order of operator **L**
* with $G(\mathbf{x}, \boldsymbol{\xi})$ considered as a function of $\mathbf{x}$, it satisfies the partial differential equation $\mathbf{L}G(\mathbf{x}, \boldsymbol{\xi}) = \delta(\mathbf{x} - \boldsymbol{\xi})$

Green's Function II

- The solution of the differential equation $LF(\mathbf{x}) = \varphi(\mathbf{x})$ is $F(\mathbf{x}) = \int_{\mathbb{R}^N} G(\mathbf{x}, \boldsymbol{\xi})\varphi(\boldsymbol{\xi})\ d\boldsymbol{\xi}$ where $\varphi(\mathbf{x})$ is a continuous function of $x \in \mathbb{R}^N$

Spectral Regularization in Fourier Domain

- A spectral operator is an integral operator $\mathbf{T}$, in the case of the Fourier operator $K(\mathbf{s}, \mathbf{x}) = exp(-j\langle \mathbf{s}, \mathbf{x} \rangle)$

Definition 8

For any functional $f(\mathbf{x}) \in H$, the Fourier operator $\mathbf{T}$ is defined by $\mathbf{T}f = F(s) = \int_{-\infty}^{+\infty} f(\mathbf{x}) exp(-j\mathbf{xs})\ d\mathbf{x}$; $F(\mathbf{s}) \in H$

Theorem 12

Given two functionals $f, g \in H$ and their corresponding Fourier transform F and G, the Parseval theorem says that $\langle f(\mathbf{x}), g(\mathbf{x}) \rangle = \frac{1}{2\pi} \langle F(\mathbf{s}), G(\mathbf{s}) \rangle$. In the operator form it is expressed by $\langle f, g \rangle = \langle \mathbf{T}f, \mathbf{T}g \rangle$

Spectral Regularization in Fourier Domain

- The differential operator $\mathbf{D} = \sum_{-\infty}^{+\infty} \frac{(-1)^n}{n!} \frac{d^n}{dx^n}$ and its spectral operator $\mathbf{T_D} = \sum_{-\infty}^{+\infty} \frac{(-1)^n (js)^n}{n!} = exp(-js)$
  - $\mathbf{T_D} f = \mathbf{T}(\mathbf{D} f) = \mathbf{D}(\mathbf{T} f)$
- The kernel function associated with the operator $\mathbf{K}$ is given by $K'(\mathbf{x}, \mathbf{x}_i) = \int exp(j\mathbf{s}\mathbf{x}) exp(-j\mathbf{s}\mathbf{x}_i)\ d\mathbf{s} = \delta(\mathbf{x} - \mathbf{x}_i)$
- Then $\mathbf{K} f = \int K'(\mathbf{x}, \mathbf{x}_i) f(\mathbf{s}) d\mathbf{s} = f(\mathbf{x} - \mathbf{x}_i)$

Regularization Strategy I

- The regularization strategy for the equation $A\varphi = f$ ($A : X \to Y$) is to find an approximated solution φ^ϵ related to φ, such that
  * $\|\varphi^\epsilon - \varphi\| \leq \epsilon$ where ϵ is a small, positive value
- The two equivalent conditions are to find:
  * $R_\lambda : Y \to X$ ($\lambda > 0$) such that $\lim_{\lambda \to 0} R_\lambda A_\varphi = \varphi$ for all $\varphi \in X$ (pointwise convergence)
  * $R_\lambda f \to A^{-1} f$ as $\lambda \to 0$
- **Note**: the operator R_λ cannot be uniformly bounded with respect to λ, and the operator $R_\lambda A$ cannot be norm convergent as $\lambda \to 0$

Regularization Strategy II

- The approximation error is

$$\|\varphi_\lambda^\epsilon - \varphi\| \le \underbrace{\epsilon\|R_\lambda\|}_{\text{influence of incorrect data}} + \underbrace{\|R_\lambda A\varphi - \varphi\|}_{\text{approximation error between } R_\lambda \text{ and } A^{-1}}$$

- The regularization parameter, λ, controls the trade-off between:
  * stability (the first term increases as $\lambda \to 0$)
  * accuracy (the second term decreases as $\lambda \to 0$)
- **Note**: the aim of the regularization is to find an appropriate λ to achieve the minimum error of the regularized risk functional

Regularization Strategy III

Definition 9

The regularization strategy R_λ i.e. the choice of regularization parameter $\lambda = \lambda(\epsilon)$, is called regular if for all $f \in A(X)$ and all $f^\epsilon \in Y$ with $\|f^\epsilon - f\| \leq \epsilon$, there holds $R_{\lambda(\epsilon)} f^\epsilon \to A^{-1} f$, $\epsilon \to 0$

Theorem 13

For a bounded linear operator $A(x)^\perp = N(A^*)$ and $N(A^*)^\perp = \overline{A(X)}$

* $A(X)^\perp$ is the orthogonal complement of $A(X)$ i.e $A^\perp = \{\varphi \in X : \langle A\varphi, g\rangle = 0 \; \forall g \in A^\perp\}$
* $N(A^*)$ is the kernel of A^* i.e. $N(A^*) = \{A^* f = 0, \; \forall f\}$

Spectral Theory I

Theorem 14

Let X be a Hilbert space, and let $A : X \to X$ be a self-adjoint compact operator

* all eigenvalues of A are real
* all eigenspaces $N(\xi I - A)$ for nonzero eigenvalues ξ have finite dimension
* all eigenspaces associated with different eigenvalues are orthogonal

Suppose the eigenvalues are ordered such that $|\xi_1| \geq |\xi_2| \geq \ldots$, and denote by $P_n : X \to N(\xi_n I - A)$ the orthogonal projection onto the eigenspace for the eigenvalue ξ_n; then

$$A = \sum_{n=1}^{\infty} \xi_n P_n$$

Let $Q : X \to N(A)$ be the orthogonal projection onto the kernel $N(A)$; then

$$\varphi = \sum_{n=1}^{\infty} P_n \varphi + Q\varphi \quad \forall \varphi \in X$$

Spectral Theory II

- In the case of an orthonormal basis, $\langle \varphi_n, \varphi_k \rangle = \delta_{n,k}$, the expansion representation is

$$A\varphi = \sum_{n=1}^{\infty} \xi_n \langle \varphi, \varphi_n \rangle \varphi_n$$

$$\varphi = \sum_{n=1}^{\infty} \langle \varphi, \varphi_n \rangle \varphi_n + Q\varphi$$

Singular Values Decomposition

Theorem 15

Let X and Y be Hilbert spaces. Let $A : X \to Y$ be a linear compact operator and $A^* : Y \to X$ be its adjoint. Let σ_n denote the singular values of A, which are the square roots of the eigenvalues of the self-adjoint compact operator $A^*A : X \to X$. Let $\{\sigma_n\}$ be an ordered sequence of nonzero singular values of A according to the dimension of the kernel $N(\sigma_n^2 I - A^*A)$. Then there exist orthonormal sequences $\{\varphi_n\}$ in X and $\{g_n\}$ in Y such that

$$A\varphi_n = \sigma_n g_n, \quad A^* g_n = \sigma_n \varphi_n \quad \forall n \in \mathbb{Z}$$

- For each $\varphi \in X$, the *Singular Values Decomposition* (SVD) is $\varphi = \sum_{n=1}^{\infty} \langle \varphi, \varphi_n \rangle \varphi_n + Q\varphi$
  - with the orthogonal projection $Q : X \to N(A)$ and $A\varphi = \sum_{n=1}^{\infty} \sigma_n \langle \varphi, \varphi_n \rangle g_n$

Picard's Theorem

Theorem 16

Let $A : X \to Y$ be a linear compact operator with singular system $(\sigma_n, \varphi_n, g_n)$. The Fredholm integral equation of the first kind $A\varphi = f$ is solvable if and only if $f \in N(A^*)^\perp$ and $\sum_{n=1}^{\infty} \frac{1}{\sigma_n^2} |\langle f, g_n \rangle|^2 < \infty$. Then a solution is given by

$$\sum_{n=1}^{\infty} \frac{1}{\sigma_n} \langle f, g_n \rangle \varphi_n$$

- The Picard theorem describes the ill-posed nature of the integral equation $A\varphi = f$
  - the perturbation ratio $\|\varphi^\epsilon\| / \|f^\epsilon\| = \epsilon / \sigma_n$ determines the degree of ill-posedness
  - more quickly σ decays and the more severe is the ill-posedness

Green's Identity

- From the Parseval identity (theorem 12), the $\mathcal{R}_{reg}(f,h)$ becomes $d\mathcal{R}_{reg}(f,h) = \int \mathbf{D}f\mathbf{D}h\ d\mathbf{x} = \langle \mathbf{D}h, \mathbf{D}f \rangle = \int \mathbf{T_D}f\overline{\mathbf{T_D}h}\ ds = \langle \mathbf{T_D}h, \mathbf{T_D}f \rangle$

Definition 10

For any pair of functions $u(\mathbf{x})$ and $v(\mathbf{x})$, given a linear differential operator $\mathbf{D}$ and its associated Fourier operator $\mathbf{T_D}$, with adjoint operators, $\mathbf{D}^*$ and $\mathbf{T_D^*}$, are uniquely determined to satisfy the boundary conditions

$$\int_{\mathbb{R}^N} u(\mathbf{x})\mathbf{D}v(\mathbf{x})d\mathbf{x} = \int_{\mathbb{R}^N} v(\mathbf{x})\mathbf{D}^*u(\mathbf{x})d\mathbf{x}$$

$$\int_{\Omega} u(\mathbf{s})\overline{\mathbf{T_D}v(\mathbf{s})}d\mathbf{s} = \int_{\Omega} v(\mathbf{s})\mathbf{T_D^*}u(\mathbf{s})d\mathbf{s}$$

where Ω represents spectrum support in the frequency domain

Euler-Lagrange equation of Tikhonov functional I

- The first equation is the Green's identity and it can be used to write

$$d\mathcal{R}_{reg}(f,h) = \langle \mathbf{D}h, \mathbf{D}f \rangle = \langle h, \mathbf{D}^*\mathbf{D}f \rangle(\mathbf{x})$$

  with $u(\mathbf{x}) = \mathbf{D}f(\mathbf{x})$ and $\mathbf{D}v(\mathbf{x}) = \mathbf{D}h(\mathbf{x})$

- From the second equation derives

$$d\mathcal{R}_{reg}(f,h) = \langle \mathbf{T_D}h, \mathbf{T_D}f \rangle = \langle h, \mathbf{T_D^*}\mathbf{T_D}f \rangle(\mathbf{s})$$

  with $u(\mathbf{s}) = \mathbf{T_D}f(\mathbf{s})$ and $\mathbf{T_D}v(\mathbf{s}) = \mathbf{T_D}h(\mathbf{s})$

- The condition $d\mathcal{R}(f,h) = d\mathcal{R}_{emp}(f,h) + \lambda d\mathcal{R}_{reg}(f,h) = 0$ becomes

$$d\mathcal{R}(f,h) = \left\langle h(\mathbf{x}), \left[\mathbf{D}^*\mathbf{D}f(\mathbf{x}) - \frac{1}{\lambda}\sum_{i=1}^{l}(y_i - f(\mathbf{x}_i))\delta(\mathbf{x}-\mathbf{x}_i)\right]\right\rangle$$

$$d\mathcal{R}(f,h) = \left\langle h(\mathbf{s}), \left[\mathbf{T_D^*}\mathbf{T_D}f(\mathbf{s}) - \underbrace{\mathcal{F}}_{\text{Fourier transform}}\left\{\frac{1}{\lambda}\sum_{i=1}^{l}(y_i - f(\mathbf{x}_i))\delta(\mathbf{x}-\mathbf{x}_i)\right\}\right]\right\rangle$$

Euler-Lagrange equation of Tikhonov functional II

- The two equivalent necessary conditions for $f(\mathbf{x})$ is an extremum of $\mathcal{R}(f)$ are
  * $d\mathcal{R}(f,h) = 0$ for all $h \in H$
  * in the distribution sense

$$\underbrace{\mathbf{D}^*\mathbf{D}f_\lambda(\mathbf{x}) = \frac{1}{\lambda}\sum_{i=1}^{l}(y_i - f(\mathbf{x}_i))\delta(\mathbf{x}-\mathbf{x}_i)}_{\text{Euler-Lagrange equation of Tikhonov functional } \mathcal{R}(f)} \qquad \text{and}$$

$$\mathbf{T}_\mathbf{D}^*\mathbf{T}_\mathbf{D}f_\lambda(\mathbf{s}) = \mathcal{F}\left\{\frac{1}{\lambda}\sum_{i=1}^{l}(y_i - f(\mathbf{x}_i))\delta(\mathbf{x}-\mathbf{x}_i)\right\} = \underbrace{\frac{1}{\lambda}\sum_{i=1}^{l}(y_i - f(\mathbf{x}_i))exp(-j\mathbf{x}_i\mathbf{s})}_{\text{Fourier transform of Euler-Lagrange equation of Tikhonov functional } \mathcal{R}(f)}$$

Differential and Integral Operators

- Consider $\mathbf{L} = \mathbf{D}^*\mathbf{D}$ and $\mathbf{K} = \mathbf{T}_\mathbf{D}^*\mathbf{T}_\mathbf{D}$
    * the $G(\mathbf{x}, \boldsymbol{\xi})$ is the Green's function for the linear differential operator $\mathbf{L}G(\mathbf{x}, \boldsymbol{\xi}) = \delta(\mathbf{x} - \boldsymbol{\xi})$
    * in frequency domain $\mathbf{K}G(\mathbf{s}, \boldsymbol{\xi}) = exp(-j\mathbf{s}\boldsymbol{\xi})$
- The operators $\mathbf{L}$ and $\mathbf{K}$ are defined as
    * $\mathbf{L} = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!2^n} \nabla^{2n}$ where $\nabla^2 = \sum_{i=1}^{l} \frac{\partial^2}{\partial x_i^2}$ is the Laplace operator
    * $\mathbf{K} = \sum_{n=0}^{\infty} \frac{(-1)^{2n} s^{2n}}{n!2^n} = exp(s^2/2)$
- Since $\mathbf{L}G(\mathbf{x}) = \delta(\mathbf{x})$ and $\mathbf{K}G(\mathbf{s}) = 1$ follow that $G(\mathbf{s}) = exp(-\|\mathbf{s}\|^2/2) \leftrightarrow G(\mathbf{x}) = exp(-\|\mathbf{x}\|^2/2)$

Solution of Regularized Problem I

- The differential equation $\mathbf{L}f(\mathbf{x}) = \varphi(\mathbf{x})$ and integral equation $\mathbf{K}f(\mathbf{x}) = \phi(\mathbf{s})$ have the same the solution $f(\mathbf{x}) = \int_{\mathbb{R}^N} G(\mathbf{x}, \boldsymbol{\xi})\varphi(\boldsymbol{\xi})\, d\boldsymbol{\xi}$
- **Note**:
  * $\mathbf{L}f(\mathbf{x}) = \mathbf{L}\int_{\mathbb{R}^N} G(\mathbf{x}, \boldsymbol{\xi})\varphi(\boldsymbol{\xi})\, d\boldsymbol{\xi} = \int_{\mathbb{R}^N} \delta(\mathbf{x} - \boldsymbol{\xi})\varphi(\boldsymbol{\xi})\, d\boldsymbol{\xi} = \varphi(\mathbf{x})$
  * $\varphi(\boldsymbol{\xi}) = \frac{1}{\lambda}\sum_{i=1}^{l}[y_i - f(\mathbf{x}_i)]\delta(\boldsymbol{\xi} - \mathbf{x}_i)$
  * $f_\lambda(\mathbf{x}) = \frac{1}{\lambda}\sum_{i=1}^{l}[y_i - f(\mathbf{x}_i)]\int_{\mathbb{R}^N} G(\mathbf{x}, \boldsymbol{\xi})\delta(\boldsymbol{\xi} - \mathbf{x}_i)\, d\boldsymbol{\xi} = \sum_{i=1}^{l} w_i G(\mathbf{x}, \mathbf{x}_i)$
  * equivalently in the frequency domain
- The purpose of regularization is to find a subspace, the eigenspace of $\mathbf{L}f$ or $\mathbf{K}f$, within which the operator becomes "well-posed". The solution in the subspace is unique

Solution of Regularized Problem II

- The solution of the regularized problem is $f_\lambda(\mathbf{x}) = \sum_{i=1}^{l} w_i G(\mathbf{x}, \mathbf{x}_i)$ with $w_i = [y_i - f(\mathbf{x}_i)]/\lambda$ and $G(\mathbf{x}, \mathbf{x}_i)$ is a positive-definite Green's function for all i
- In general, the solution of the Tikhonov regularization problem is $f_\lambda(\mathbf{x}) = \sum_{i=1}^{l} w_i G(\mathbf{x}, \mathbf{x}_i) + \beta(\mathbf{x})$ where
  * $\beta(x)$ is a term that lies in the kernel of $\mathbf{D}$ which satisfies the orthogonal condition $\sum_{i=1}^{l} w_i \beta(\mathbf{x}_i) = 0$
  * the simplest case is $\beta(x) = const$
  * the functional space of the solution f_λ is an RKHS of the direct sum of two orthogonal RKHS

Parzen's Window

- Replacing the summation by the integral and $G(\mathbf{x}, \mathbf{x}_i)$ by $K(\mathbf{x}, \mathbf{x}_i)$. If K is a reproducing kernel invariant to translation and satisfies $K(\mathbf{x}, \mathbf{x}_i) = K(\mathbf{x} - \mathbf{x}_i)$. The approximated function can be expressed by the convolution

$$f(\mathbf{x}) = \frac{1}{\lambda} \int \underbrace{(y_i - f(\mathbf{x}_i))}_{\text{observation}} \underbrace{K(\mathbf{x} - \mathbf{x}_i)}_{\text{kernel}} \, d\mathbf{x}_i$$

$$= \frac{1}{\lambda} \Big(\underbrace{\int y_i K(\mathbf{x} - \mathbf{x}_i) \, d\mathbf{x}_i}_{\text{Kernel regression}} - \underbrace{\int f(\mathbf{x}_i) K(\mathbf{x} - \mathbf{x}_i) \, d\mathbf{x}_i}_{\text{Integral operator}} \Big)$$

- The $f(\mathbf{x})$ can be reconstructed by the data sample smoothed by an averaged kernel (Parzen's Window)

Regularized Solution in Matrix Form

- The regularized solution in matrix form is obtained by taking $\frac{d\mathcal{R}(f)}{df}$ and setting it to zero

$$\mathcal{R}(f) = \frac{1}{2}\|\mathbf{y} - \mathbf{f}\|^2 + \frac{1}{2}\lambda(\mathbf{D}f)^T(\mathbf{D}f) = \frac{1}{2}(\mathbf{y} - \mathbf{f})^T(\mathbf{y} - \mathbf{f}) + \frac{1}{2}\lambda\mathbf{f}^T \underbrace{K}_{\mathbf{D}^*\mathbf{D}} \mathbf{f}$$

where the solution is $\mathbf{f} = \underbrace{(\mathbf{I} - \lambda K)^{-1}}_{\text{smoothing matrix}} \mathbf{y}$

  * K is quadratic symmetric penalty matrix associated with $\mathbf{L}$
  * similarly in the frequency domain where the regularizer is $\mathbf{K}$

References

Andreas Kirsch.
An Introdutction to the Mathematical Theory of the Inverse Problems - Second Edition.
Springer, 2011.

S. Haykin Z. Chen.
On Different Facets of Regularization Theory.
Neural Computation, 14:2791–2846, 2002.