A Different Type of Convergence for Statistical Learning Algorithms

Cynthia Rudin

PACM, Fine Hall, Princeton University, Princeton NJ 08544, USA,
crudin@princeton.edu,
WWW home page: http://www.math.princeton.edu/~crudin

Abstract. We discuss stability for a class of learning algorithms with respect to noisy labels. The algorithms we consider are for regression, and they involve the minimization of regularized risk functionals, such as $L(f) := \frac{1}{N}\sum_{i=1}^{N}(f(\mathbf{x}_i) - y_i)^2 + \lambda\|f\|_{\mathcal{H}}^2$. We shall call the algorithm 'stable' if, when y_i is a noisy version of $\overline{f}(\mathbf{x}_i)$ for some function $\overline{f} \in \mathcal{H}$, the output of the algorithm converges to $\overline{f}$ as the regularization term and noise simultaneously vanish. We consider two flavors of this problem, one where a data set of N points remains fixed, and the other where $N \to \infty$. For the case where $N \to \infty$, we give conditions for convergence to $f_{\mathbb{E}}$ (the function which is the expectation of $y(\mathbf{x})$ for each $\mathbf{x}$), as $\lambda \to 0$. For the fixed N case, we describe the limiting 'non-noisy', 'non-regularized' function $\overline{f}$, and give conditions for convergence. In the process, we develop a set of tools for dealing with functionals such as $L(f)$, which are applicable to many other problems in learning theory.

keywords statistical learning theory, learning in the limit, regularized least squares regression, RKHS
Note Please consider me for the student travel support and for the Mark Fulk award. Thank you.

1 Introduction

In regression learning problems, we are given data $(\mathbf{x}_i, y_i)_{i=1,..,N}$ in $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}$ is a bounded subset of $\mathbb{R}^n$ and $\mathcal{Y}$ is a bounded subset of $\mathbb{R}$. We assume this data is chosen iid (independently and identically distributed) according to an unknown probability distribution $\mu(\mathbf{x}, y)$. We say that $\mathbf{x}$ is a 'position', and y is a 'label'. These data points may be, for example, images of people's faces in pixel space with a person's age as the corresponding label, or auto-regressive time series data ([4], [6]). The output of a learning algorithm is a decision function $f : \mathcal{X} \to \mathbb{R}$. Even though we only know N data points from distribution $\mu(\mathbf{x}, y)$, we hope to construct f which will be able to *generalize* to unobserved points in the distribution. This means we would like f to predict the value of y for any given value of $\mathbf{x} \in \mathcal{X}$. Since we want our function f to fit the data accurately and also have this generalization ability, we refer to Vapnik's Structural Risk Minimization (SRM) principle ([10], [11]). In SRM, we limit our choice of functions f so they are chosen from a class $\mathcal{F}$, of finite 'capacity' (i.e. finite VC dimension). Otherwise, we cannot hope to choose a function f which has generalization ability -

we would overfit the data. One convenient way to implement SRM is to let $\mathcal{F}$ be a ball within a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$, with norm $\|\cdot\|_{\mathcal{H}}$. In this form, we have an Ivanov regularization problem; one can show that the solution is always the minimizer of a corresponding Tikhonov regularization problem. Algorithms for classification and regression solve this Tikhonov Regularization problem, so that the decision function is given by f_{min} ([2], [8]), where

$$f_{min} := \operatorname*{argmin}_f L(f), \text{ where } L(f) = \frac{1}{N}\sum_{i=1}^{N} V(f(\mathbf{x}_i) - y_i) + \lambda \|f\|_{\mathcal{H}}^2 \ .$$

$L(f)$ is called the Regularized Risk functional. Note that we define our RKHS $\mathcal{H}$ so that $f \in \mathcal{H}$ iff $\|f\|_{\mathcal{H}}$ is finite. Thus, minimizing over $f \in \mathcal{H}$ is equivalent to minimizing over all functions f. The first term in $L(f)$ is called the Empirical Risk, and $V(\cdot)$ is a predetermined loss function. We will generally use the least squares loss function $V(z) = (z)^2$, but a similar analysis can be performed for other loss functions. The second term is called the Regularization term, and λ is called the Regularization parameter; one always takes $\lambda > 0$. Here, λ can be viewed as the trade-off between accuracy and generalization. If λ is very small, we are minimizing the Empirical Risk, increasing the accuracy of our model to the data, and possibly overfitting. If λ is very large, our algorithm will generalize at the expense of accuracy. In a sense, λ controls the capacity of the function class from which f is chosen: the larger λ is, the smaller the radius of the ball $\mathcal{F}$ in $\mathcal{H}$. In practice, λ is often chosen empirically, perhaps to minimize the leave-one-out error on a training set.

Another interpretation of this functional is through the eyes of algorithmic stability, as described by Bousquet and Elisseeff ([1]). Here, the regularization term prevents the algorithm from being sensitive to the replacement of one data point. In either case, the regularization problem is well-posed only when λ is strictly greater than 0.

We assume that the labels y are 'noisy', in the sense that there is a marginal distribution $\mu(y|\mathbf{x})$ for each $\mathbf{x}$. We denote the expectation value of the label y for position x as $\mathbb{E}(y|\mathbf{x})$, and we denote the marginal distribution along the $\mathbf{x}$-axis as $\mu(\mathbf{x})$. (This is the distribution of $\mu(\mathbf{x}, y)$ after integrating over the y values.)

For the case when $N \to \infty$, we show convergence of f_{min} to a function $f_{\mathbb{E}}$ as the regularization term vanishes, provided $f_{\mathbb{E}} \in \mathcal{H}$; i.e. we need to find conditions on the simultaneous convergence $N \to \infty$ and $\lambda \to 0$ so that $f_{min} \to f_{\mathbb{E}}$. Here, the function $f_{\mathbb{E}}$ is defined by:

$$f_{\mathbb{E}} = \operatorname*{argmin}_f \textit{Actual Risk}\,(f), \text{ where } \textit{Actual Risk}\,(f) = \int \left(\mathbb{E}[f(\mathbf{x}) - y]^2 | \mathbf{x}\right) d\mu(\mathbf{x}).$$

In other words, $f_{\mathbb{E}}$ is the minimizer of the *'Actual Risk'*. Since we are using the least squares loss function, this minimizer is simply the expectation of y for each $\mathbf{x}$; $f_{\mathbb{E}}(\mathbf{x}) = \mathbb{E}(y|\mathbf{x})$.

We assume that we have chosen a RKHS which is large enough to contain $f_{\mathbb{E}}(\mathbf{x})$. In other words, $\|f_{\mathbb{E}}(\mathbf{x})\|_{\mathcal{H}} < \infty$. This is not an exceedingly strong assumption; in fact, many popular kernels (e.g. gaussian kernels) can produce RKHS of arbitrarily high VC dimension. Although $f_{\mathbb{E}}(\mathbf{x})$ may not be in $\mathcal{H}$ for every case, $f_{\mathbb{E}}(\mathbf{x})$ will be in $\mathcal{H}$ for most

smooth processes which have bounded noise, as long as we implement a sufficiently powerful RKHS.

For the fixed N case, we may express label y for position $\mathbf{x}$ as the random variable $y(\mathbf{x}) = \tilde{f}(\mathbf{x}) + b(\mathbf{x})$, where $\tilde{f} : \mathcal{X} \to \mathcal{Y}$ is a deterministic function assumed to be in $\mathcal{H}$, and $b(\mathbf{x})$ is random noise with some probability distribution, with $b(\mathbf{x})$ and $b(\mathbf{x}')$ independent if $\mathbf{x} \neq \mathbf{x}'$. We denote the vector of noise values as $\mathbf{b} = (b_1, b_2, .., b_N) = \{b(\mathbf{x}_i)\}_{i=1,..,N}$. In order to force the noise to vanish, we will assume the noise is generated by a fixed random process generating noise with norm bounded by b_{max} almost surely, and we will only shrink its amplitude. Since the noise is generated by this fixed process, the theorem will hold whenever the noise is bounded, and thus, if the noise is bounded almost surely, the theorem will hold almost surely. Using the least squared loss and making the noise explicit, our algorithm becomes:

$$f_{min} := \underset{f}{\operatorname{argmin}}\ L(f), \text{ where } L(f) = \frac{1}{N}\sum_{i=1}^{N}(f(\mathbf{x}_i) - \tilde{f}(\mathbf{x}_i) - c\, b_i)^2 + \lambda\|f\|_H^2,$$
$$\text{where } \|\mathbf{b}\|_{\ell_2} \leq \sqrt{N} b_{max} \text{ and } c \text{ is a constant.}$$

For $\lambda > 0$, the minimizer of $L(f)$ is unique, because $\lambda\|f\|_{\mathcal{H}}^2$ is strictly convex. Since the noise is random, f_{min} is still a random variable. Our goal is to show 'stability' for this algorithm, i.e. we need to find a set of conditions on the simultaneous convergence $\lambda, c \to 0$ which allows $f_{min} \to \overline{f}$, where $\overline{f}$ is the element of $\mathcal{H}$ with minimal norm that has zero Empirical Risk when noise is not present. (Since we assume that $\tilde{f} \in \mathcal{H}$, $\tilde{f}$ itself minimizes $L(f)$ when $\lambda = 0$. Since many functions in $\mathcal{H}$ vanish at all the $\mathbf{x}_i$, there may be infinitely many functions with zero empirical risk; our algorithm will converge to the one with the smallest RKHS norm.)

Intuitively, this stability analysis demonstrates that there's no inherent error in our algorithm when noise or regularization is present, and that a small amount of noise or regularization cannot dramatically disrupt the algorithm's output. This type of stability is different from the 'algorithmic stability' of Devroye([3]). Algorithmic stability measures the variability of an algorithm's output as the data set changes. Our type of stability determines whether the algorithm's output changes dramatically when noise or regularization is present. Algorithmic stability is a property of one particular algorithm for one particular distribution of data. Our stability is not - the distribution changes as noise is removed, and the algorithm changes as the regularization term shrinks. We actually use algorithmic stability to help us show stability of our algorithm in this sense.

Theorem 1 states that the regularized least squares regression algorithm is stable as the number of data points increases to infinity. Theorem 2 states that the regularized least squares regression algorithm is stable for a fixed N point data set.

Main Algorithm (Regularized Least Squares Regression):
For a data set $Z = (\mathbf{x}_i, y_i)_{i=1,..,N}$, where $\forall i \in 1, .., N$, $\mathbf{x}_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$

$$f_{Z,\lambda} := \underset{f}{\operatorname{argmin}}\ L_{Z,\lambda}(f), \text{ where } L_{Z,\lambda}(f) = \frac{1}{N}\sum_{i=1}^{N}(f(\mathbf{x}_i) - y_i)^2 + \lambda\|f\|_{\mathcal{H}}^2\ .$$

Theorem 1. *Denote by $f_{\mathbb{E}}(\mathbf{x})$ the function $\mathbb{E}(y|\mathbf{x})$. Denote by Z_N the data set $(\mathbf{x}_i, y_i)_{i=1,..,N}$. If $f_{\mathbb{E}} \in \mathcal{H}$ and if $\lambda := \lambda_N$ is chosen to depend on N such that $\lambda_N \to 0$ and $N\lambda_N^3 \to \infty$ as $N \to \infty$, then we have convergence of the Main Algorithm:*

$$\|f_{Z_N,\lambda_N} - f_{\mathbb{E}}\|_{\mathcal{H}} \xrightarrow{P} 0 \text{ as } N \to \infty.$$

Here, '$\xrightarrow{P}$' denotes convergence in probability.

Theorem 2. *Assume we are given N fixed positions $\mathbf{x}_1, .., \mathbf{x}_N$.*
Suppose that for each $i \in 1, .., N$, the labels are given by $y_i = \tilde{f}(\mathbf{x}_i) + \frac{1}{t}b_i$, where the b_i's are independent random variables with $\|\mathbf{b}\|_{\ell_2} \leq \sqrt{N}b_{max}$ almost surely. Denote by Z_t the data set $(\mathbf{x}_i, y_i)_{i=1,..,N}$.

Define $\overline{f}$ by:

(i) *$\overline{f}(\mathbf{x}_i) = \tilde{f}(\mathbf{x}_i)$ for $i = 1, .., N$*

(ii) *$\|\overline{f}\|_{\mathcal{H}}$ is minimal, among all functions which satisfy **(i)**.*

If $\lambda := \lambda_t$ is chosen to depend on t such that $t\sqrt{\lambda_t} \to \infty$ as $t \to \infty$ and $\lambda_t \to 0$, then we have convergence of the Main Algorithm almost surely:

$$\|f_{Z_t,\lambda_t} - \overline{f}\|_{\mathcal{H}} \to 0 \text{ as } t \to \infty \quad \text{almost surely .}$$

Section 2 contains a short review of RKHS. Section 3 and 4 contain the proofs of Theorems 1 and 2.

2 Reproducing Kernel Hilbert Space (RKHS)

$\mathcal{H}$ is a real Reproducing Kernel Hilbert Space (RKHS) if $\mathcal{H}$ has the following properties:

- *$\mathcal{H}$ is a Hilbert space.* $\mathcal{H}$ is a complete, inner product, real vector space of functions $f : \mathcal{X} \to \mathbb{R}$. We denote $\mathcal{H}$'s inner product by $(\cdot, \cdot)_{\mathcal{H}}$, and $\mathcal{H}$'s norm by $\|\cdot\|_{\mathcal{H}}$.
- *Reproducing Property.* There exists a bilinear form $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $\forall \mathbf{x} \in \mathcal{X}$, we have $K(\mathbf{x}, \cdot) \in \mathcal{H}$ and $(f, K(\mathbf{x}, \cdot))_{\mathcal{H}} = f(\mathbf{x})$ for any $f \in \mathcal{F}$. This K is called the 'reproducing kernel' of the RKHS.[8][9][2]. We sometimes denote $K(\mathbf{x}, \cdot)$ by $K_{\mathbf{x}}$.
- *Spanning Property.* $\mathcal{H} = \overline{span\{K(\mathbf{x}, \cdot)|\mathbf{x} \in \mathcal{X}\}}$

Since $\mathcal{H}$ is a real Hilbert space, $(f, g) = (g, f)$ for all $f, g \in \mathcal{H}$. It follows that $K(\mathbf{x}, \mathbf{x}') = (K(\mathbf{x}, \cdot), K(\mathbf{x}', \cdot))_{\mathcal{H}} = (K(\mathbf{x}', \cdot), K(\mathbf{x}, \cdot))_{\mathcal{H}} = K(\mathbf{x}', \mathbf{x})$, i.e. K is symmetric in its two arguments. An equivalent definition of an RKHS is a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$ such that all evaluation functionals $\Gamma_{\mathbf{x}} : f \to f(\mathbf{x}),\ \mathbf{x} \in \mathcal{X}$, are continuous. Given $\mathbf{x}_1, .., \mathbf{x}_N \in \mathcal{X}$, the associated $N \times N$ Gram Matrix G has entries $G_{ij} = K(x_i, x_j)$ where K is the reproducing kernel for the RKHS $\mathcal{H}$. The Gram Matrix is always a positive semi-definite matrix.

The Representer Theorem transforms the minimization of our functional $L_{Z,\lambda}(f)$ into an optimization problem over only N numbers. This advantage is the main reason why scientists take $\mathcal{F}$ to be a ball in a RKHS $\mathcal{H}$. We present a corollary of this theorem below.

Corollary of the Representer Theorem (Kimeldorf, Wahba)[5]) *The function* $f_{min} = argmin_f \ \frac{1}{N}\sum_{i=1}^{N} V(f(\mathbf{x}) - y_i) + \lambda\|f\|_{\mathcal{H}}^2$ *can be represented in the form* $f_{min} = \sum_{i=1}^{N} \alpha_i K_{x_i}$. *This is true for any arbitrary loss function V.*
(This corollary is a specific case of the full Representer Theorem [5].)

Having described the basic facts about RKHS, we now continue with the proofs of Theorems 1 and 2.

3 Proof of Theorem 1

For the *Main Algorithm* above, we are increasing the size of the data set Z_N as N increases. We need to show convergence $f_{Z_N,\lambda_N} \xrightarrow{P} f_{\mathbb{E}}$, where $\lambda_N \to 0$ and $N \to \infty$. That is, we need to show

$$\lim_{N\to\infty} P\{\|f_{Z_N,\lambda_N} - f_{\mathbb{E}}\|_{\mathcal{H}} \geq \eta\} = 0 \text{ for every } \eta > 0 \ .$$

We can break up the distance $\|f_{Z_N,\lambda_N} - f_{\mathbb{E}}\|_{\mathcal{H}}$ into two contributions. The first contribution is called 'variance', and it is due to the finite number of randomly chosen noisy data points. The variance vanishes with arbitrarily high probability as the number of data points increases, even if the noise does not vanish. The second contribution is the 'bias' due to the restriction we place on our hypothesis space, i.e. the fact that f is chosen from with a ball of a RKHS. This term vanishes as the ball gets larger, i.e. when λ_N gets smaller.

$$\|f_{Z_N,\lambda_N} - f_{\mathbb{E}}\|_{\mathcal{H}} \leq \underbrace{\|f_{Z_N,\lambda_N} - \hat{f}_{\lambda_N}\|_{\mathcal{H}}}_{\text{variance}} + \underbrace{\|\hat{f}_{\lambda_N} - f_{\mathbb{E}}\|_{\mathcal{H}}}_{\text{bias}}$$

where $f_{Z_N,\lambda_N} = \underset{f}{\text{argmin}}\ L_{Z_N,\lambda_N}(f)$,

$$\text{where } L_{Z_N,\lambda_N}(f) = \tfrac{1}{N}\textstyle\sum_{i=1}^{N}(f(\mathbf{x}_i) - y_i)^2 + \lambda_N\|f\|_H^2, \text{ and}$$

$\hat{f}_{\lambda_N} = \underset{f}{\text{argmin}}\ \hat{L}_{\lambda_N}(f)$, where $\hat{L}_{\lambda_N}(f) = \int(\mathbb{E}[f(\mathbf{x}) - y]^2|\mathbf{x})d\mu(\mathbf{x}) + \lambda_N\|f\|_H^2$.

Lemma 1.1 below describes a method for proving that the minimizers of two convex functions are close in $\mathcal{H}$.

Lemma 1.1. *Suppose* $L^1, L^2 : \mathcal{H} \to \mathbb{R}$ *are two convex functionals for which there exist* ε,δ *so that:*

(a) $\forall\ f \in \mathcal{H}\ ;\ |L^1(f) - L^2(f)| < \varepsilon$

(b) $|L^1(f) - L^1(f^1)| < 2\varepsilon \Longrightarrow \|f - f^1\|_{\mathcal{H}} < \delta$

Then if the minimizers $f^1 := \underset{f}{\text{argmin}}\ L^1(f)$ *and* $f^2 := \underset{f}{\text{argmin}}\ L^2(f)$ *exist, they satisfy* $\|f^1 - f^2\|_{\mathcal{H}} < \delta$.

Proof. Since f^1 and f^2 are minimizers of L^1 and L^2 respectively, and using the closeness condition **(a)**:

$$\begin{gathered} L^1(f^1) \leq L^1(f^2) \leq L^2(f^2) + \varepsilon \\ L^2(f^2) \leq L^2(f^1) \leq L^1(f^1) + \varepsilon \\ \text{So, } |L^1(f^1) - L^2(f^2)| \leq \varepsilon. \end{gathered}$$

Now, $|L^1(f^1) - L^1(f^2)| \leq |L^1(f^1) - L^2(f^2)| + |L^2(f^2) - L^1(f^2)| \leq 2\varepsilon$, and finally, condition **(b)** will give us $\|f^1 - f^2\|_{\mathcal{H}} \leq \delta$. □

Back to the proof of Theorem 1. We proceed one term at a time.

Variance Term We will choose more general versions of $L_{Z_N,\lambda_N}(f)$ and $\hat{L}_{\lambda_N}(f)$ temporarily.

$$\begin{aligned} \mathcal{L}_{Z_N,\lambda_N}(f) &:= \frac{1}{N}\sum_{i=1}^{N} V(f(\mathbf{x}_i) - y_i) + \lambda_N \|f\|_{\mathcal{H}}^2 \\ \hat{\mathcal{L}}_{\lambda_N}(f) &:= \int \mathbb{E}(V(f(\mathbf{x}) - y)|\mathbf{x}) d\mu(\mathbf{x}) + \lambda_N \|f\|_{\mathcal{H}}^2 \\ &\text{where } |V(a) - V(b)| \leq \sigma_V |a - b|. \end{aligned}$$

That is, we assume that the loss function V is Lipschitz continuous, or 'sigma-admissible' [1]. The least squares loss has $\sigma_V = 2\mathcal{X}_{max}$, since $|V(a) - V(b)| = |a^2 - b^2| \leq |a + b||a - b| \leq 2\mathcal{X}_{max}|a - b|$.

We need to verify the conditions **(a)** and **(b)** in order to use Lemma 1.1. To verify the closeness property **(a)** for our functionals $\mathcal{L}_{Z_N,\lambda_N}(f)$ and $\hat{\mathcal{L}}_{\lambda_N}(f)$:

$$\begin{aligned} \left|\mathcal{L}_{Z_N,\lambda_N}(f) - \hat{\mathcal{L}}_{\lambda_N}(f)\right| &= \left|\frac{1}{N}\sum_{i=1}^{N} V(f(\mathbf{x}_i) - y_i) - \int \mathbb{E}(V(f(\mathbf{x}) - y)|\mathbf{x}) d\mu(\mathbf{x})\right| \\ &= \Big|\ \textit{Empirical Risk}\ (f)\ -\ \textit{Actual Risk}\ (f)\ \Big| \end{aligned} \tag{1}$$

There are many available upper bounds for the right side of equation (1), including Vapnik's VC bound, which relies on the VC dimension of the class of allowed decision functions $\mathcal{F}$ ([10], [11]). The particular bound we utilize for this paper was constructed by Bousquet and Elisseeff [1], and it is based on 'algorithmic stability'. In general, bounds of this quantity are probabilistic, and are based on some capacity measure of the algorithm or space of functions $\mathcal{F}$. This particular bound relies on the sigma-admissibility of the loss function V, and McDiarmid's concentration of measure inequality.

def $\mathcal{Z}_N$ is a training sample $(\mathbf{x}_1, y_1), .., (\mathbf{x}_N, y_N)$. $\mathcal{Z}^i_{N;\underline{\mathbf{x}},\underline{y}} = (\mathcal{Z}_N \backslash (\mathbf{x}_i, y_i)) \cup (\underline{\mathbf{x}}, \underline{y})$. That is, we replace the i^{th} training point in $\mathcal{Z}_N$ by a new data point in order to obtain $\mathcal{Z}^i_{N;\underline{\mathbf{x}},\underline{y}}$.

def The algorithm $Alg : \mathcal{Z} \to f_{\mathcal{Z}}$ is *uniformly β-stable with respect to loss function* $V_\beta : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ if: $|V_\beta(\mathbf{x}, y, f_{\mathcal{Z}_N}(\mathbf{x})) - V_\beta(\mathbf{x}, y, f_{\mathcal{Z}^i_{N;\underline{\mathbf{x}},\underline{y}}}(\mathbf{x}))| \leq \beta$ for all $(\mathbf{x}, y),\ (\underline{\mathbf{x}}, \underline{y}) \in \mathcal{X} \times \mathcal{Y}, i,$ and all $\mathcal{Z}_N$.

Basically, this algorithmic stability measures how much the algorithm's output could possibly change, as measured by the loss function V_β, when we replace one data point.

Algorithmic Stability Theorem (Bousquet and Elisseeff, [1]) *If we are given a uniformly β-stable algorithm with respect to loss function V_β, which outputs functions bounded by the constant M (i.e. $|f_Z(\mathbf{x})| \leq M \ \ \forall\, \mathbf{x} \in \mathcal{X},\ \forall \mathcal{Z}$), then for any $N \geq \frac{8M^2}{\varepsilon^2}$,*

$$P\{|\textit{Empirical Risk}(f) - \textit{Actual Risk}(f)| > \varepsilon\} \leq p_N, \text{with } p_N = \frac{64MN\beta + 8M^2}{N\varepsilon^2}.$$

Algorithmic Stability of Tikhonov Learning Algorithms (Bousquet and Elisseeff, [1]) *The Main Algorithm is uniformly β-stable, with $\beta = \frac{C^2\kappa^2}{N\lambda_N}$. Here, C is an upper bound on the sigma-admissibility constant σ_V, and κ is an upper bound on the diagonal elements of the Gram Matrix, that is, $max_i\ \ G_{ii} \leq \kappa$.*

Returning to the proof of Theorem 1, we now know the right side of (1) is bounded by ε (for large values of N) with probability at least $1 - p_N$, where:

$$p_N = \frac{64MN\beta + 8M^2}{N\varepsilon^2}, \text{ where } \beta = \frac{C^2\kappa^2}{N\lambda_N}, \text{ so } p_N = \frac{64MC^2\kappa^2 + 8M^2\lambda_N}{N\lambda_N\varepsilon^2}.$$

We now have the closeness condition **(a)** of Lemma 1.1 satisfied with probability at least $1 - p_N$, i.e. $|L_{Z_N,\lambda_N}(f) - \hat{L}_{\lambda_N}(f)| \leq \varepsilon$, with probability at least $1 - p_N$, where p_N is given in (2).

Now we verify condition **(b)**. We need to show that $|\hat{L}_{\lambda_N}(f) - \hat{L}_{\lambda_N}(\hat{f}_{\lambda_N})| \leq 2\varepsilon$ implies that $\|f - \hat{f}_{\lambda_N}\|_{\mathcal{H}} \leq \delta$ for every f. Let's define a function h so that $h := f - \hat{f}_{\lambda_N}$ in what follows.

$$\begin{aligned}
\hat{L}_{\lambda_N}(f) &= \int \mathbb{E}(\hat{f}_{\lambda_N}(\mathbf{x}) + h(\mathbf{x}) - y \,|\, \mathbf{x})^2 d\mu(\mathbf{x}) + \lambda_N \|\hat{f}_{\lambda_N} + h\|^2_{\mathcal{H}} \\
&= \hat{L}_{\lambda_N}(\hat{f}_{\lambda_N}) + \Big[2 \int \mathbb{E}[(\hat{f}_{\lambda_N}(\mathbf{x}) - y)h(\mathbf{x})|\mathbf{x}] d\mu(\mathbf{x}) + \ 2\lambda_N(\hat{f}_{\lambda_N}, h)_{\mathcal{H}}\Big] \\
&\quad + \int h^2(\mathbf{x}) d\mu(\mathbf{x}) + \lambda_N \|h\|^2_{\mathcal{H}}
\end{aligned}$$

The terms in the brackets are linear in h. Remember that $\hat{f}_{\lambda_N}$ is the minimizer of $\hat{L}_{\lambda_N}$, and thus the linear terms must be zero. (If the linear terms are non-zero, we can reverse

the sign of h and contradict $\hat{f}_{\lambda_N}$ as the minimizer.) The last two terms are always positive.

$$\hat{L}_{\lambda_N}(f) = \hat{L}_{\lambda_N}(\hat{f}_{\lambda_N}) + \int h^2(\mathbf{x})d\mu(\mathbf{x}) + \lambda_N\|h\|_{\mathcal{H}}^2 \geq \hat{L}_{\lambda_N}(\hat{f}_{\lambda_N}) + \lambda_N\|h\|_{\mathcal{H}}^2$$

Now we can see that **(b)** holds:

$$\|f - \hat{f}_{\lambda_N}\|_{\mathcal{H}}^2 = \|h\|_{\mathcal{H}}^2 \leq (\hat{L}_{\lambda_N}(f) - \hat{L}_{\lambda_N}(\hat{f}_{\lambda_N}))\frac{1}{\lambda_N} \leq \frac{2\varepsilon}{\lambda_N}$$

In this case,

$$\delta = \sqrt{\frac{2\varepsilon}{\lambda_N}}\,. \tag{2}$$

Since both the conditions **(a)** and **(b)** are satisfied, Lemma 1.1 produces

$$\|f_{Z_N,\lambda_N} - \hat{f}_{\lambda_N}\|_{\mathcal{H}} \leq \sqrt{\frac{2\varepsilon}{\lambda_N}} \text{ with prob. at least } 1 - p_N,$$

where p_N is given in (2). We are done with the variance term.

Bias Term We will prove that the bias term vanishes using the spectral theorem. Define the function $h(\mathbf{x}) := f(\mathbf{x}) - f_{\mathbb{E}}(\mathbf{x})$ in what follows. Now,

$$\hat{L}_{\lambda_N}(f) = \int (K_{\mathbf{x}}, h)_{\mathcal{H}}^2 d\mu(\mathbf{x}) + \int \mathbb{E}[(f_{\mathbb{E}}(\mathbf{x}) - y)^2|\mathbf{x}]d\mu(\mathbf{x}) + \lambda_N\|h + f_{\mathbb{E}}\|_{\mathcal{H}}^2$$

The minimizer of $\hat{L}_{\lambda_N}(f)$ again must have first variational derivative equal to 0. Using Fubini's Theorem, we find:

$$\begin{aligned}\frac{\partial \hat{L}_{\lambda_N}(f + \gamma g)}{\partial \gamma}\bigg|_{\gamma=0} &= \frac{\partial}{\partial \gamma}\left[\int (K_{\mathbf{x}}, h + \gamma g)_{\mathcal{H}}^2 d\mu(\mathbf{x}) + \lambda_N\|h + f_{\mathbb{E}} + \gamma g\|_{\mathcal{H}}^2\right]\bigg|_{\gamma=0} \\ &= 2\left(\int (K_{\mathbf{x}}, h)K_{\mathbf{x}}d\mu(\mathbf{x}) + \lambda_N(h + f_{\mathbb{E}}), g\right)_{\mathcal{H}}\end{aligned}$$

If $f = \hat{f}_{\lambda_N}$, the above expression must be zero for all g, thus:

$$0 = \int (K_{\mathbf{x}}, \hat{f}_{\lambda_N} - f_{\mathbb{E}})_{\mathcal{H}}K_{\mathbf{x}}d\mu(\mathbf{x}) + \lambda_N(\hat{f}_{\lambda_N}) \tag{3}$$

Let's define a new operator T.
def $T: \mathcal{H} \longrightarrow \mathcal{H}$
$f \longmapsto \int (K_{\mathbf{x}}, f)_{\mathcal{H}}K_{\mathbf{x}}d\mu(\mathbf{x})$

One can check that T is self-adjoint since $(Tf, g)_{\mathcal{H}} = \int (K_{\mathbf{x}}, f)_{\mathcal{H}}(K_{\mathbf{x}}, g)_{\mathcal{H}}d\mu(\mathbf{x}) = (f, Tg)_{\mathcal{H}}$. For an operator Q from one Hilbert space H_1, to another, H_2, the operator

norm of Q is defined by $\|Q\|_{\mathcal{L}(H_1,H_2)} := \sup_{\|s\|_{H_1}=1} \|Qs\|_{H_2}$. Our operator T is bounded, since by Cauchy-Schwarz,

$$\|T\|^2_{\mathcal{L}(\mathcal{H},\mathcal{H})} \leq \sup_{\|f\|_{\mathcal{H}}=1} \|f\|^2_{\mathcal{H}} \int\int \sqrt{K(\mathbf{x},\mathbf{x})}K(\mathbf{x},\mathbf{z})\sqrt{K(\mathbf{z},\mathbf{z})}d\mu(\mathbf{x})d\mu(\mathbf{z}) \;\leq\; \infty.$$

We are going to use the spectral theorem next, but first let us review a few facts from functional analysis about this theorem ([7]). The spectral theorem allows one to define functions of a bounded self-adjoint operator on a Hilbert space H. If the function is a polynomial, e.g. $f(z) = 3x^2 - 5z + 2$, then it is clear how to define the corresponding operator $\phi(A) : \phi(A) = 3A^2 - 5A + 2$. The spectral theorem extends the correspondence $\phi(z) \leftrightarrow \phi(A)$ to all continuous functions (in fact, to all bounded Borel functions). Moreover, one has $\|\phi(A)\|_{\mathcal{L}(H,H)} \leq sup\{|\phi(z)| \;:\; z \in spec(A)\}$. Because ϕ is a real function, the operator $\phi(A)$ provided by the spectral theorem is also self-adjoint. In addition, for each $f \in H$, we have a measure $\nu_{f;A}$ on $spec(A)$ such that $(\phi(A)f, f)_H = \int_{spec(A)} \phi(z)d\nu_{f;A}(z)$. The measure $\nu_{f;A}$ is concentrated on that part of the spectrum $spec(A)$ along which f has a nonzero component. In particular, if $f \in Ker(A)$, then f is an eigenvector of A with eigenvalue 0, and $\nu_{f;A}$ is a δ-measure concentrated on $\{0\}$. If on the contrary, $f \perp Ker(A)$, then $\nu_{f;A}(\{0\}) = 0$.

Now, using the definition of the spectral measure $\nu_{f_{\mathbb{E}};T}$ for the operator T and the function $f_{\mathbb{E}}$, we have from (3):

$$\begin{aligned} 0 &= T(\hat{f}_{\lambda_N} - f_{\mathbb{E}}) + \lambda_N(\hat{f}_{\lambda_N}) \\ \Rightarrow\quad \hat{f}_{\lambda_N} - f_{\mathbb{E}} &= (T + \lambda_N)^{-1}(-\lambda_N f_{\mathbb{E}}) \\ \Rightarrow\quad \|\hat{f}_{\lambda_N} - f_{\mathbb{E}}\|^2_{\mathcal{H}} &= \|(T + \lambda_N)^{-1}\lambda_N f_{\mathbb{E}}\|^2_{\mathcal{H}} = \int_{spec(T)} \left(\frac{\lambda_N}{\gamma + \lambda_N}\right)^2 d\nu_{f_{\mathbb{E}};T}(\gamma) \end{aligned}$$

Since $K(\cdot,\cdot)$ is positive semidefinite, T is a positive operator and thus has non-negative spectrum only. One can see that $Ker\ T$ is empty, i.e. take any function ϑ such that $T\vartheta = 0$. Then, $0 = (T\vartheta, \vartheta)_{\mathcal{H}} = \int \vartheta^2(\mathbf{x})d\mu(\mathbf{x})$; thus, ϑ must be zero almost everywhere. It follows that $\{0\}$ is a set of measure zero for $\nu_{f_{\mathbb{E}};T}$. As $N \to \infty$, $\lambda_N \to 0$, and the function $(\frac{\lambda_N}{\gamma+\lambda_N})$ converges to 0 pointwise on $\mathbb{R}_+$, and thus almost everywhere with respect to $\nu_{f_{\mathbb{E}};T}$; since this function is bounded by 1, we can again use the dominated convergence theorem to say that the integral vanishes as $N \to \infty$. One cannot give a more explicit bound for this term without more information about the relationship between μ and $\mathcal{H}$. In any case, we have convergence of the bias term to 0.

Now, we can complete the proof of Theorem 1. For any $\eta > 0$, we must be able to show that $\lim_{N\to\infty} P\{\|f_{Z_N,\lambda_N} - f_{\mathbb{E}}\|_{\mathcal{H}} \geq \eta\} = 0$. So, let us choose an arbitrary fixed value η. The bias term vanishes as $N \to \infty$ and $\lambda_N \to 0$, so there must exist an N_0 so that for $N > N_0$, λ_N is sufficiently small, so that the bias term is bounded by $\eta/2$. Thus, we consider the bias term bounded by $\eta/2$ in the limit as $N \to \infty$; since this term does not depend on the data, the bound clearly holds with probability 1. Now, we must choose ε_N so that the variance term is bounded by $\eta/2$. Using the bound in (2), we choose $\varepsilon_N = \frac{\eta^2\lambda_N}{8}$. The corresponding probability p_N is then given by (2),

$$p_N = \frac{64MC^2\kappa^2 + 8M^2\lambda_N}{N\lambda_N\varepsilon_N^2} = \frac{64(64MC^2\kappa^2 + 8M^2\lambda_N)}{N\lambda_N^3\eta^4}.$$

We need p_N to vanish as $N \to \infty$; this is satisfied if $N\lambda_N^3 \to \infty$ as $N \to \infty$. Also, there must exist an N_0 such that for $N > N_0$, we have $N \geq \frac{8M^2}{\varepsilon_N^2}$; we need this in order to use the Algorithmic Stability Theorem. Thus, Theorem 1 is proved. □

4 Proof of Theorem 2

This section contains the proof of Theorem 2. First, some notation. The positions $\mathbf{x}_1, .., \mathbf{x}_N$ will be considered fixed throughout this section.

def $P : \mathcal{H} \longrightarrow \mathbb{R}^N$
$f \longmapsto (f(\mathbf{x}_1), f(\mathbf{x}_2), .., f(\mathbf{x}_N))$

The 'evaluation operator' P evaluates a function f at each position $\mathbf{x}_i$ in the data set. Note that P 'loses information' about a function f by evaluating it at only N points. That is, $Ker\ P$ is a nontrivial subspace of $\mathcal{H}$. The adjoint $P^* : \mathbb{R}^N \longrightarrow \mathcal{H}$ of the operator P is given by $P^* : (c_1, .., c_N) \longmapsto \sum_{i=1}^N c_i K_{\mathbf{x}_i}$. One can show that P is a bounded operator, with $\|P\|_{\mathcal{L}(H,H)} = \|P^*P\|^{1/2}_{\mathcal{L}(\mathcal{H},\mathcal{H})} \leq (\max_j \sum_{i=1}^N |G_{ij}|)^{1/2}$. The operator P^*P is automatically positive and self-adjoint. We will later use the spectral theorem on the bounded self-adjoint operator P^*P.

We start the proof of Theorem 2 with the following Lemma.

Lemma 2.1: *The following characterizations of $\overline{f}$ are equivalent*:

1. $\overline{f}$ satisfies:
   (i) $\overline{f}(\mathbf{x}_i) = \tilde{f}(\mathbf{x}_i)$ for $i = 1, .., N$, and
   (ii) $\|\overline{f}\|_{\mathcal{H}} \leq \|g\|_{\mathcal{H}}\ \ \forall g \in \mathcal{H}$ that satisfy $g(\mathbf{x}_i) = \tilde{f}(\mathbf{x}_i)$.
2. $\overline{f}$ satisfies:
$$\overline{f} = \sum_{i=1}^N \tilde{f}(\mathbf{x}_i) W_{\mathbf{x}_i}, \text{ where } W_{\mathbf{x}_i} = \sum_{\ell=1}^N G^{-1}_{i\ell} K_{\mathbf{x}_\ell} \tag{4}$$
3. $\overline{f}$ satisfies:
   (i) $\overline{f}(\mathbf{x}_i) = \tilde{f}(\mathbf{x}_i)$ for $i = 1, .., N$, and
   (iii) $\forall h \in Ker\ P$ we have $(\overline{f}, h)_{\mathcal{H}} = 0$.

Proof. We will show 1. → 2. → 3. → 1.
1. ↔ 2. First, we show that the function described in 1. is unique. From the reproducing property, we know that $\overline{f}$ has nonzero components along each of the $K_{\mathbf{x}_i}$'s for which $\overline{f}(\mathbf{x}_i) \neq 0$. Since $\mathcal{H}$ is a Hilbert space, we can always decompose $\overline{f}$ into a component $f_\parallel$ within the span of the $K_{\mathbf{x}_i}$'s and a component $f_\perp$ orthogonal to each $K_{\mathbf{x}_i}$ (where $i \in 1, .., N$). Now, $\|\overline{f}\|^2_{\mathcal{H}} = \|f_\parallel\|^2_{\mathcal{H}} + \|f_\perp\|^2_{\mathcal{H}} \geq \|f_\parallel\|^2_{\mathcal{H}}$. Thus, if $\|f_\perp\|_{\mathcal{H}} \neq 0$, then $\overline{f}$ no longer has minimal norm and contradicts property **(ii)**. The component of $\overline{f}$ along each

of the $K_{\mathbf{x}_i}$'s is determined by the value of $\overline{f}(\mathbf{x}_i)$. So, functions f that satisfy both **(i)** and **(ii)** can be written $f = f_{\|} = \sum_{i=1}^{N} \alpha_i K_{\mathbf{x}_i}$ for the fixed values of α_i, $i = 1, .., N$. In particular, the α_i's must satisfy:

$$\sum_{i=1}^{N} \alpha_i G_{ij} = \sum_{i=1}^{N} \alpha_i K_{\mathbf{x}_i}(\mathbf{x}_j) = f(\mathbf{x}_j) \ .$$

Thus, the function described in 1. is unique. It is straightforward to see that the function described in 2. is exactly the function described in 1. Evaluating the right side of (4) at $\mathbf{x}_j$, we obtain

$$\overline{f}(\mathbf{x}_i) = \sum_{j=1}^{N} \tilde{f}(\mathbf{x}_j) \sum_{\ell=1}^{N} G_{j\ell}^{-1} G_{\ell i} = \tilde{f}(\mathbf{x}_i).$$

Moreover, the function described in 2. lies entirely within the span of the $K_{\mathbf{x}_i}$'s. Therefore it obeys **(i)** and **(ii)** and we have 1. $\leftrightarrow$ 2.

2. $\rightarrow$ 3. Because 2. $\rightarrow$ 1., **(i)** is satisfied. We just need to show **(iii)**.
For any $h \in Ker\ P$,

$h(\mathbf{x}_\ell) = 0 \quad \forall\, \ell \in 1, .., N$
$(h, K_{\mathbf{x}_\ell})_{\mathcal{H}} = 0 \quad \forall\, \ell \in 1, .., N$ by the reproducing property,
$(h, W_{\mathbf{x}_i})_{\mathcal{H}} = 0 \quad \forall\, \ell, i \in 1, .., N$ because the $W_{\mathbf{x}_i}$'s are each a linear combination of the $K_{\mathbf{x}_\ell}$'s.
$(h, \overline{f})_{\mathcal{H}} = 0$ because $\overline{f}$ is a linear combination of the $W_{\mathbf{x}_i}$'s, thus **(iii)** holds.

3. $\rightarrow$ 1. Here, **(i)** is automatic, so we need to check **(ii)**.

Take arbitrary $g \in \mathcal{H}$ with: $g(\mathbf{x}_i) = \tilde{f}(\mathbf{x}_i) = \overline{f}(\mathbf{x}_i)$ for i=1,..,N.
Then, $g - \overline{f} \in Ker\ P$.
From assumption **(iii)**, $(\overline{f}, g - \overline{f})_{\mathcal{H}} = 0$,
and thus $(\overline{f}, g)_{\mathcal{H}} = \|\overline{f}\|_{\mathcal{H}}^2$.

Now,

$$\begin{aligned} \|g\|_H^2 &= \|g - \overline{f} + \overline{f}\|_{\mathcal{H}}^2 \\ &= \|g - \overline{f}\|_{\mathcal{H}}^2 + \|\overline{f}\|_{\mathcal{H}}^2 + 2(g - \overline{f}, \overline{f})_{\mathcal{H}} \\ &= \|g - \overline{f}\|_{\mathcal{H}}^2 + \|\overline{f}\|_{\mathcal{H}}^2 \\ &\geq \|\overline{f}\|_{\mathcal{H}}^2, \quad \text{with equality only if } g = \overline{f}. \end{aligned}$$

Thus we have 1. $\rightarrow$ 2. $\rightarrow$ 3. $\rightarrow$ 1., so Lemma 2.1 is proved. □

Back to the proof of Theorem 2. The functional $L_{Z_t,\lambda_t}(f)$ in the *Main Algorithm* expressed in terms of P becomes:

$$L_{Z_t,\lambda_t}(f) = \frac{1}{N} \left\| Pf - (P\tilde{f} + \frac{1}{t}\mathbf{b}) \right\|_{\ell_2}^2 + \lambda_t \|f\|_{\mathcal{H}}^2$$

The minimizer of $L_{Z_t,\lambda_t}(f)$ must satisfy $\frac{\partial}{\partial\gamma}L_{Z_t,\lambda_t}(f+\gamma h)\big|_{\gamma=0}=0$. In other words, the first variational derivative of $L_{Z_t,\lambda_t}(f)$ is 0 at its minimizer. Recalling that $P\tilde{f}=P\overline{f}$, this minimization problem becomes:

$$\begin{aligned}0&=\frac{\partial}{\partial\gamma}\big[\frac{1}{N}(Pf+\gamma Ph-P\overline{f}-\frac{1}{t}\mathbf{b},Pf+\gamma Ph-P\overline{f}-\frac{1}{t}\mathbf{b})_{\ell_2}+\lambda_t\|f+\gamma h\|^2_{\mathcal{H}}\big]\Big|_{\gamma=0}\\&=2\frac{1}{N}(Ph,Pf-P\overline{f}-\frac{1}{t}\mathbf{b})_{\ell_2}+2\lambda_t(f,h)_{\mathcal{H}}\\&=2\Big(h,\frac{1}{N}P^*Pf-\frac{1}{N}P^*P\overline{f}-\frac{1}{N}P^*\Big(\frac{1}{t}\mathbf{b}\Big)+\lambda_t f\Big)_{\mathcal{H}}.\end{aligned}$$

This must be true for any function h, so $\frac{1}{N}P^*Pf-\frac{1}{N}P^*P\overline{f}-\frac{1}{N}P^*\Big(\frac{1}{t}\mathbf{b}\Big)+\lambda_t f=0$, implying

$$f=\overline{f}-(\frac{1}{N}P^*P+\lambda_t)^{-1}\lambda_t\overline{f}+(\frac{1}{N}P^*P+\lambda_t)^{-1}\frac{1}{N}P^*\Big(\frac{1}{t}\mathbf{b}\Big)\,.$$

It follows that

$$\|f-\overline{f}\|_{\mathcal{H}}\le\|(\frac{1}{N}P^*P+\lambda_t)^{-1}\lambda_t\overline{f}\|_{\mathcal{H}}+\frac{1}{Nt}\|(\frac{1}{N}P^*P+\lambda_t)^{-1}P^*\|_{\mathcal{L}(\ell_2,\mathcal{H})}\|\mathbf{b}\|_{\ell_2}.$$

In order to show stability, we bound the two terms on the right of (5), and construct these bounds so they vanish as $t\to\infty$. That is, we need to bound the norms above. To accomplish this, we will use the spectral theorem on the bounded self-adjoint operator P^*P.

To bound the first term in equation (5), recall that the operator obtained from the function $\phi_t(z)=(\frac{1}{N}z+\lambda_t)^{-2}\lambda_t^2$ of the self-adjoint operator P^*P is self-adjoint. Also, since P^*P is a positive operator, the spectrum $spec(P^*P)$ of the operator P^*P is concentrated on $\mathbb{R}_+\cup\{0\}$. Using the spectral measure $\nu_{\overline{f};P^*P}(z)$ on the spectrum $spec(P^*P)$, we find:

$$\begin{aligned}\|(\frac{1}{N}P^*P+\lambda_t)^{-1}\lambda_t\overline{f}\|^2_{\mathcal{H}}&=([(\frac{1}{N}P^*P+\lambda_t)^{-1}\lambda_t]^2\overline{f},\overline{f})_{\mathcal{H}}.\\&=\int_{spec(P^*P)}\Big(\frac{\lambda_t}{\frac{1}{N}z+\lambda_t}\Big)^2\,d\nu_{\overline{f};P^*P}(z)\end{aligned}$$

Because $\lambda_t\underset{t\to\infty}{\longrightarrow}0$, we have $\phi_t(z)=(\frac{1}{N}z+\lambda_t)^{-2}\lambda_t^2\underset{t\to\infty}{\longrightarrow}0$ for all $z\in\mathbb{R}_+\setminus\{0\}$. By Lemma 2.1, we know that $\overline{f}\perp Ker\ P$, and since $Ker\ P=Ker\ P^*P$, we know that $\nu_{\overline{f};P^*P}(\{0\})=0$. Since $\phi_t(z)\le1$ for all $z\in spec(P^*P)\subset\mathbb{R}_+$, it then follows from the dominated convergence theorem that $\|(\frac{1}{N}P^*P+\lambda_t)^{-1}\lambda_t\overline{f}\|^2_{\mathcal{H}}\underset{t\to\infty}{\longrightarrow}0$. One cannot give a more explicit bound for this first term; it would require more specific knowledge of the relationship between μ and $\mathcal{H}$. In any case, we have achieved our goal in showing that the first term of (5) vanishes as $t\to\infty$.

We need the second term in equation (5) to vanish also. Recall that for operator $Q:H_1\to H_2$, it is true that $\|Q^*\|_{\mathcal{L}(H_2,H_1)}=\|Q^*Q\|^{1/2}_{\mathcal{L}(H_1,H_1)}$, and that a continuous

real function of a self adjoint operator such as P^*P is self adjoint.

$$\|(\frac{1}{N}P^*P+\lambda_t)^{-1}P^*\|_{\mathcal{L}(\ell_2,\mathcal{H})}=\|(\frac{1}{N}P^*P+\lambda_t)^{-1}P^*P(\frac{1}{N}P^*P+\lambda_t)^{-1}\|^{1/2}_{\mathcal{L}(\mathcal{H},\mathcal{H})}.$$

We use the spectral theorem for the bounded self-adjoint operator $A=P^*P$, namely the fact $\|\phi(A)\|_{\mathcal{L}(H,H)} \leq sup\{|\phi(z)|; z \in spec(A)\}$ where $\phi(z)=\frac{z}{(\frac{1}{N}z+\lambda_t)^2}$ here. The maximum value of $\phi(z)$ occurs at $z=N\lambda_t$, and it is $\frac{N}{4\lambda_t}$. Thus,

$$\|(\frac{1}{N}P^*P+\lambda_t)^{-1}P^*\|_{\mathcal{L}(\ell_2,\mathcal{H})} \leq \frac{\sqrt{N}}{2\sqrt{\lambda_t}}$$

$$\frac{1}{Nt}\|(\frac{1}{N}P^*P+\lambda_t)^{-1}P^*\|_{\mathcal{L}(\ell_2,\mathcal{H})}\|\mathbf{b}\|_{\ell_2} \leq \frac{\sqrt{N}}{2Nt\sqrt{\lambda_t}}\|\mathbf{b}\|_{\ell_2} \leq \frac{1}{2t\sqrt{\lambda_t}}b^{max} \text{ a.s.}$$

As long as we design λ_t so that $t\sqrt{\lambda_t} \underset{t\to\infty}{\longrightarrow} \infty$, then we have the desired convergence of this term to 0. We are done with the second term of equation (5). Theorem 2 is proven. □

5 Conclusion

We have proved stability for the regularized least squares regression algorithm, for the sense in which inverse problems are examined. We have shown stability for this algorithm in two cases: the case when the number of data points N is a constant, and the case where $N \to \infty$. It is important that our algorithm is stable in this sense, because we do not want any inherent error in the algorithm's output. Neither a small amount of noise in the data nor a small amount of regularization should drastically influence the algorithm's output.

We hope that the reader will gain more from our result than the knowledge that regularized least squares regression is stable in the inverse operator sense. We have found the particular methods introduced in the proofs of Theorem 1 and Theorem 2 useful for various learning problems, especially those which require the convexity of learning functionals or convergence of learning algorithms. Namely, we demonstrate two methods for showing that the minimizers of two learning functionals are close: use of the spectral theorem, and the technique of Lemma 1.1, which can both be generally applied to other learning algorithms.

6 Acknowledgements

The author would like to express infinite gratitude and millions of thank-you's to her advisor, Dr. Ingrid Daubechies.

References

1. O. Bousquet and A. Elisseeff. Algorithmic stability and generalization performance. In Advances in Neural Information Processing Systems 13: Proc. NIPS'2000, 2001.

2. F. Cucker and S. Smale. On the mathematical foundations of learning. Bulletin (New Series) of the American Mathematical Society, 39(1):1,49, 2002.
3. L. P. Devroye and T. J. Wagner. Distribution-free performance bounds for potential function rules. IEEE Trans. Inform. Theory, 25(5):601-604, 1979.
4. B. Heisele, A. Verri, and T. Poggio, Learning and Vision Machines. IEEE Visual Perception: Technology and Tools, 90(7):1164-1177, 2002.
5. G.S. Kimeldorf and G. Wahba. Some Results on Tchebycheffian spline functions. Journal of Mathematical Analysis and Applications, 33:82-85, 1971.
6. S. Mukherjee, E. Osuna, and F. Girosi. Nonlinear Prediction of Chaotic Time Series Using Support Vector Machines, IEEE Workshop on Neural Networks for Signal Processing VII, 1997.
7. M. Reed, and B. Simon, Methods of Modern Mathematical Physics I: Functional Analysis. Academic Press, San Diego, CA, 1980.
8. B. Schoelkopf, and A. Smola, Learning with Kernels - Support Vector Machines, Regularization, Optimization, and Beyond. The MIT Press, Cambridge, MA. 2002.
9. A.N. Tikhonov and V.Y.Arsenin. Solution of Ill-Posed Problems. Winston, Washington, DC, 1977.
10. V. Vapnik. Statistical Learning Theory. John Wiley and Sons Inc., New York, 1998. A Wiley-Interscience Publication.
11. V. Vapnik. The Nature of Statistical Learning Theory. Springer Verlag, New York, 1995.