Uniform Stability of Regularized Kernel Model

Ziyang Gong *

Date: January 9, 2022

1 Introduction

Suppose $\mathcal{X}$ be the input space, $\mathcal{Y}$ be the output space, $\mathcal{D}$ be some (almost) completely unknown probability distribution on $\mathcal{X} \times \mathcal{Y}$. Given the n i.i.d observed data, which sampled from an unknown distribution $\mathcal{D}$, that,

$$S := \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} \subseteq \mathcal{D} \tag{1}$$

and the goal of us is to estimate the functional relationship between $\mathcal{X}$ and $\mathcal{Y}$.

To formalize the problem, we now aim at finding a predictor fucntion f^* among the function space $\mathcal{F} := \{f : \mathcal{X} \to \mathcal{Y}\}$ based on the observed data S, which minimizes the true risk

$$R[f] := \mathbb{E}_{\mathcal{D}}\left[L\left(y, f(\mathbf{x})\right)\right] \tag{2}$$

where $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is an arbitrary convex loss function, typically assumed that the smaller $L\left(y, f(\mathbf{x})\right)$ is, the better the approximmation of y is. Thus, we are trying to find a predictor f^* with risk close to the optimal risk

$$R^* := \inf\left\{R[f] \mid f : \mathcal{X} \to \mathcal{Y}\right\} \tag{3}$$

Finding the predictor function f^* which minimizing the empirical risk

$$R_{\text{emp}}(f) = \frac{1}{n}\sum_{i=1}^{n} L\left(y_i, f(\mathbf{x}_i)\right) \tag{4}$$

is a natural things for us to be trying to do. However, as it known to all, just minimizing the empirical risk is suicidal, which almost certainly leads to overfitting. Minimizing R_{emp} only makes sense if we simultaneously somehow restrict ourselves to the $\mathcal{F}$, that are of just the right level of complexity. One way to do this is by explicitly restricting the function space $\mathcal{F}$ to a "simple" space, as in structural risk minimization, which is to introduce a penalty functional $\Omega[f]$ that somehow measures the complexity of each function $f \in \mathcal{F}$, and to minimize the regularized risk

$$R_{\text{reg}}[f] = R_{\text{emp}}[f] + \Omega[f] \tag{5}$$

*Email: meetziyang@outlook.com

In this report, we restrict the predictor function $f \in \mathcal{F}$ among the reproducing kernel Hilbert space $\mathcal{H}$, and the regularized risk has the form

$$R_{\text{reg}}[f] = \frac{1}{n}\sum_{i=1}^{n} L\left(y_i, f(\mathbf{x}_i)\right) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2 \tag{6}$$

thus, we can estimate f^* by soving the following optimization problem

$$\hat{f} = \arg\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} L\left(y_i, f(\mathbf{x}_i)\right) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2 \tag{7}$$

where $\lambda > 0$ is a regularization parameter to reduce the danger of overfitting. Since $L\left(y, f(\mathbf{x})\right)$ is convex in f, the minimizer $\hat{f}$ is uniquely determined and a simple gradient descent algorithm can be used to find $\hat{f}$. So the main focus of this reports is to answer a remain question we want to know, whether the risk $R[\hat{f}]$ is close to the optimal risk R^*, which will influence the stability of our algorithm.

2 Some Notations and Concepts

Before getting into the formal discussion, we will introduce some notations and concepts.

- $S^{\setminus i} := S \setminus \{(\mathbf{x}_i, y_i)\}$ be the sample where the i-th observation is removed.
- $S^i := S^{\setminus i} \cup \{(\mathbf{x}, y)\}$ be the sample where the i-th observation is replaced by $(\mathbf{x}, y)$.

and let $\hat{f}_{\setminus i}$ be the estimated result based on sample $S^{\setminus i}$, $\hat{f}_i$ based on sample S^i and $\hat{f}$ based on sample S.

In order to quantify the stability of our algorithm, we will introduce one important concepts — **Uniform Stability**.

Definition 2.1 (Uniform Stability)**.** The algorithm is uniformly β-stable with respect to the loss function $L\left(y, f(\mathbf{x})\right)$, if for all samples $S := \{\mathbf{x}_i, y_i\}_{i=1}^n \subset \mathcal{D}$ and $i \in [n]$,

$$\sup_{(x,y)\in\mathcal{D}} \left| L\left(y, \hat{f}(\mathbf{x})\right) - L\left(y, \hat{f}_{\setminus i}(\mathbf{x})\right)\right| \leq \beta \tag{8}$$

i.e. the algorithm is "stable" with respect to removing a single sample at all points.

3 Uniform Stability of Regularized Kernel Model

Firstly, we will provide an auxiliary lemma.

Lemma 3.1 (Convex Functions and Derivatives)**.** *For any differentiable convex function* $f : \mathbb{R} \to \mathbb{R}$ *and any* $a, b \in \mathbb{R}$*, we have*

$$\left[f'(a) - f'(b)\right](a - b) \geq 0 \tag{9}$$

Proof. Due to the convexity of f we know that $f(a) + (b-a)f'(a) \leq f(b)$ and, likewise, $f(b) + (a-b)f'(b) \leq f(a)$. Summing up both inequalities and subtracting the terms in $f(a)$ and $f(b)$ proves (9). □

Then, we will show that the algorithm we studied in this paper satisfied definition 2.1, and the corresponding value of β can be calculated.

Theorem 3.2 (Algorithmic Stability of Risk Minimizers Bousquet and Elisseeff (2002); Schölkopf and Smola (2001))**.** *The algorithm that minimizing the regularized empirical risk in* (6) *has stability*

$$\beta = \frac{2C^2\kappa^2}{n\lambda} \tag{10}$$

where κ is a bound on $\|k(x,\cdot)\| = \sqrt{k(x,x)}$, $\|\cdot\|$ is the RKHS norm induced by k, and C is a bound on the Lipschitz constant of the loss function $L(y, f(\mathbf{x}))$, which can be viewed as a function of f.

Remark. We can see that the stability of the algorithm depends on the regularization constant via $\frac{1}{\lambda n}$, hence we may be able to afford to choose weaker regularization if the sample size n increases.

Proof. In order to distinguish between different training sets, we use $R_{\text{reg}}[f, S]$ and $R_{\text{reg}}[f, S^{\setminus i}]$ (and likewise $R_{\text{emp}}[f, S]$) during the remainder of the proof.

Since $\hat{f}$ minimizes $R_{\text{reg}}[f, S]$, that is, the **functional derivative** (Stéphane Canu, 2014) of $R_{\text{reg}}[f, S]$ at $\hat{f}$ vanishes, and so does $R_{\text{reg}}[f, S^{\setminus i}]$ at $\hat{f}_{\setminus i}$,

$$\begin{aligned} \partial_f R_{\text{reg}}\left[\hat{f}, S\right] &= \partial_f R_{\text{emp}}\left[\hat{f}, S\right] + \lambda\hat{f} = 0 \\ \partial_f R_{\text{reg}}\left[\hat{f}_{\setminus i}, S^{\setminus i}\right] &= \partial_f R_{\text{emp}}\left[\hat{f}_{\setminus i}, S^{\setminus i}\right] + \lambda\hat{f}_{\setminus i} = 0 \end{aligned} \tag{11}$$

Next, we construct an auxiliary risk function $\tilde{R}[f]$ by

$$\tilde{R}[f] := \left\langle \partial_f R_{\text{emp}}\left[\hat{f}, S\right] - \partial_f R_{\text{emp}}\left[\hat{f}_{\setminus i}, S^{\setminus i}\right], f - \hat{f}_{\setminus i} \right\rangle + \frac{\lambda}{2}\left\| f - \hat{f}_{\setminus i} \right\|_{\mathcal{H}}^2 \tag{12}$$

Clearly $\tilde{R}[f]$ is a convex function in f (the first term is linear, the second quadratic).

Additionally, by construction, we have

$$\tilde{R}[\hat{f}_{\setminus i}] = 0 \tag{13}$$

Furthermore, taking the functional derivative of $\tilde{R}[f]$, that,

$$\partial_f \tilde{R}[f] = \partial_f R_{\text{emp}}\left[\hat{f}, S\right] - \partial_f R_{\text{emp}}\left[\hat{f}_{\setminus i}, S^{\setminus i}\right] + \lambda\left(f - \hat{f}_{\setminus i}\right) = \partial_f R_{\text{emp}}\left[\hat{f}, S\right] + \lambda f \tag{14}$$

the functional derivative of $\tilde{R}[f]$ (14) vanishes at $f = \hat{f}$ due to (11), thus the minimum of $\tilde{R}[f]$ is obtained for $f = \hat{f}$. Therefore, combined with $\tilde{R}[\hat{f}_{\setminus i}] = 0$, we can conclude that $\tilde{R}[\hat{f}] \leq 0$.

In order to obtain bounds on $\left\|\hat{f}-\hat{f}_{\setminus i}\right\|$, we have to get rid of some of the first terms in $\tilde{R}[f]$, since

$$\begin{aligned}
&n\left\langle\partial_f R_{\text{emp}}\left[\hat{f},S\right]-\partial_f R_{\text{emp}}\left[\hat{f}_{\setminus i},S^{\setminus i}\right],\hat{f}-\hat{f}_{\setminus i}\right\rangle\\
=&\sum_{j\neq i}\left[L'\left(y_j,\hat{f}(\mathbf{x}_j)\right)-L'\left(y_j,\hat{f}_{\setminus i}(\mathbf{x}_j)\right)\right]\left[\hat{f}(\mathbf{x}_j)-\hat{f}_{\setminus i}(\mathbf{x}_j)\right]\\
&+L'\left(y_i,\hat{f}(\mathbf{x}_i)\right)\left[\hat{f}(\mathbf{x}_i)-\hat{f}_{\setminus i}(\mathbf{x}_i)\right]\\
\geq&L'\left(y_i,\hat{f}(\mathbf{x}_i)\right)\left[\hat{f}(\mathbf{x}_i)-\hat{f}_{\setminus i}(\mathbf{x}_i)\right]
\end{aligned}\tag{15}$$

The first equation is due to the fact that the functional derivative $\partial_f(f)=k(\mathbf{x},.)$ and then collecting the common terms between $R_{\text{emp}}\left[\hat{f},S\right]$ and $R_{\text{emp}}\left[\hat{f}_{\setminus i},S^{\setminus i}\right]$. And, as for the last inequation, we use lemma 3.1 applied to the loss function $L\left(y,f(\mathbf{x})\right)$ which is a convex function of $f(\mathbf{x})$.

Combine the above result with the fact $\tilde{R}[\hat{f}]\leq 0$, we have

$$\left\langle\partial_f R_{\text{emp}}\left[\hat{f},S\right]-\partial_f R_{\text{emp}}\left[\hat{f}_{\setminus i},S^{\setminus i}\right],\hat{f}-\hat{f}_{\setminus i}\right\rangle+\frac{\lambda}{2}\left\|\hat{f}-\hat{f}_{\setminus i}\right\|_{\mathcal{H}}^2\leq 0\tag{16}$$

thus,

$$L'\left(y_i,\hat{f}(\mathbf{x}_i)\right)\left[\hat{f}(\mathbf{x}_i)-\hat{f}_{\setminus i}(\mathbf{x}_i)\right]+\frac{n\lambda}{2}\left\|\hat{f}-\hat{f}_{\setminus i}\right\|_{\mathcal{H}}^2\leq 0\tag{17}$$

and by the convexity of loss function $L\left(y,f(\mathbf{x})\right)$,

$$\begin{aligned}
\frac{n\lambda}{2}\left\|\hat{f}-\hat{f}_{\setminus i}\right\|_{\mathcal{H}}^2\leq&L'\left(y_i,\hat{f}(\mathbf{x}_i)\right)\left[\hat{f}_{\setminus i}(\mathbf{x}_i)-\hat{f}(\mathbf{x}_i)\right]\\
\leq&L\left(y_i,\hat{f}(\mathbf{x}_i)\right)-L\left(y_i,\hat{f}_{\setminus i}(\mathbf{x}_i)\right)\\
\leq&\left|L\left(y_i,\hat{f}(\mathbf{x}_i)\right)-L\left(y_i,\hat{f}_{\setminus i}(\mathbf{x}_i)\right)\right|
\end{aligned}\tag{18}$$

By the Cauchy-Schwarz inequality we can see that, for any $f,f'\in\mathcal{H}$ and any $\mathbf{x}\in\mathcal{X}$,

$$|f(\mathbf{x})-f'(\mathbf{x})|=|\langle f-f',k(\mathbf{x},\cdot)\rangle|\leq\|f-f'\|_{\mathcal{H}}\|k(\mathbf{x},\cdot)\|_{\mathcal{H}}\leq\kappa\|f-f'\|_{\mathcal{H}}\tag{19}$$

and since $L\left(y,f(\mathbf{x})\right)$ is Lipschitz continuous at $\mathbf{x}_i$, we have

$$\left|L\left(y,\hat{f}(\mathbf{x}_i)\right)-L\left(y,\hat{f}_{\setminus i}(\mathbf{x}_i)\right)\right|\leq C\left|\hat{f}(\mathbf{x}_i)-\hat{f}_{\setminus i}(\mathbf{x}_i)\right|\leq C\kappa\left\|\hat{f}-\hat{f}_{\setminus i}\right\|_{\mathcal{H}}\tag{20}$$

Combine equation (18) and (20), we get

$$\frac{n\lambda}{2}\left\|\hat{f}-\hat{f}_{\setminus i}\right\|_{\mathcal{H}}^2\leq C\kappa\left\|\hat{f}-\hat{f}_{\setminus i}\right\|_{\mathcal{H}}\tag{21}$$

thus,

$$\left\|\hat{f}-\hat{f}_{\setminus i}\right\|_{\mathcal{H}}\leq\frac{2C\kappa}{n\lambda}\tag{22}$$

Therefore, by the equation (20) for every $\mathbf{x}$, we have

$$\left|L\left(y,\hat{f}(\mathbf{x})\right)-L\left(y,\hat{f}_{\setminus i}(\mathbf{x})\right)\right|\leq C\kappa\left\|\hat{f}-\hat{f}_{\setminus i}\right\|_{\mathcal{H}}\leq\frac{2C^2\kappa^2}{n\lambda}\tag{23}$$

□

Within the uniform stability of our algorithm, we will also prove that the β-stable algorithm exhibit uniform convergence of the empirical risk $R_{\text{emp}}[f]$ to the true risk $R[f]$.

Theorem 3.3 (McDiarmid's Bound (McDiarmid, 1989)). *Suppose $\xi_1, \ldots, \xi_n$ be i.i.d real value random variables and assume that there exists a function $g : \mathbb{R}^n \to \mathbb{R}$ with the property that for all $i \in [n]$ and $c_i > 0$,*

$$\sup_{\xi_1,\ldots,\xi_n,\xi_i' \in \xi} |g(\xi_1, \ldots, \xi_n) - g(\xi_1, \ldots, \xi_{i-1}, \xi_i', \xi_{i+1}, \ldots, \xi_n)| \leq c_i \tag{24}$$

where ξ_i' is drawn from the same distribution as ξ_i. Then

$$\mathbb{P}\{|g(\xi_1, \ldots, \xi_n) - \mathbb{E}[g(\xi_1, \ldots, \xi_n)]| > \varepsilon\} \leq 2\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right) \tag{25}$$

Theorem 3.4 (Bousquet and Elisseeff (Bousquet and Elisseeff, 2001; Ofer Dekel and Thach Nguyen, 2011)). *Assume that we have a β-stable algorithm with the additional requirement that the loss function $L(y, f(\mathbf{x})) \leq M$ for all $(\mathbf{x}, y) \in \mathcal{D}$ and for all samples $S := \{\mathbf{x}_i, y_i\}_{i=1}^n \subset \mathcal{D}$. Then, for any $n \geq 1$*

$$\mathbb{P}\left\{\left|R_{emp}[\hat{f}, S] - R[\hat{f}]\right| > \varepsilon + 2\beta\right\} \leq 2\exp\left(-\frac{n\varepsilon^2}{2(n\beta + M)^2}\right) \tag{26}$$

Proof. Within the i.i.d assumption, we have

$$\mathbb{E}_{S\sim\mathcal{D}}\left[R_{\text{emp}}[\hat{f}]\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{S\sim\mathcal{D}}\left[L\left(y_i, \hat{f}(\mathbf{x}_i)\right)\right] = \mathbb{E}_{S\sim\mathcal{D}}\left[L\left(y_i, \hat{f}(\mathbf{x}_i)\right)\right] \tag{27}$$

If we replace $(\mathbf{x}_i, y_i)$ by $(\mathbf{x}, y)$, we can get

$$\mathbb{E}_{S\sim\mathcal{D}}\left[R_{\text{emp}}[\hat{f}]\right] = \mathbb{E}_{S,(\mathbf{x},y)\sim\mathcal{D}}\left[L\left(y, \hat{f}_i(\mathbf{x})\right)\right] \tag{28}$$

and with the observation that

$$\mathbb{E}_{\mathcal{D}}\left[R[\hat{f}]\right] = \mathbb{E}_{S,(\mathbf{x},y)\sim\mathcal{D}}\left[L\left(y, \hat{f}(\mathbf{x})\right)\right] \tag{29}$$

In order to bound on the expected difference between $R_{\text{emp}}[\hat{f}, S]$ and $R[\hat{f}]$, which leads to

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[R_{\text{emp}}[\hat{f}, S] - R[\hat{f}]\right] =& \mathbb{E}_{S,(\mathbf{x},y)\sim\mathcal{D}}\left[L\left(y, \hat{f}_i(\mathbf{x})\right)\right] - \mathbb{E}_{S,(\mathbf{x},y)\sim\mathcal{D}}\left[L\left(y, \hat{f}(\mathbf{x})\right)\right] \\ =& \mathbb{E}_{S,(\mathbf{x},y)\sim\mathcal{D}}\left[L\left(y, \hat{f}_i(\mathbf{x})\right) - L\left(y, \hat{f}(\mathbf{x})\right)\right] \leq 2\beta\end{aligned} \tag{30}$$

By the triangle inequality, we have

$$\left|R[\hat{f}] - R[\hat{f}_i]\right| \leq \left|R[\hat{f}] - R[\hat{f}_{\setminus i}]\right| + \left|R[\hat{f}_{\setminus i}] - R[\hat{f}_i]\right| \leq 2\beta \tag{31}$$

Also, we have

$$\begin{aligned}\left|R_{\text{emp}}[\hat{f}, S] - R_{\text{emp}}\left(\hat{f}_i, S^i\right)\right| \leq& \frac{1}{n}\sum_{j\neq i}\left|L\left(y_j, \hat{f}(\mathbf{x}_j)\right) - L\left(y_j, \hat{f}_i(\mathbf{x}_j)\right)\right| \\ &+ \frac{1}{n}\left|L\left(y_i, \hat{f}(\mathbf{x}_i)\right) - L\left(y_i, \hat{f}_i(\mathbf{x}_i)\right)\right| \\ \leq& \frac{n-1}{n}2\beta + \frac{2M}{n} \leq 2\beta + \frac{2M}{n}\end{aligned} \tag{32}$$

and,

$$\begin{aligned}\left|\left[R_{\text{emp}}[\hat{f}, S] - R[\hat{f}]\right] - \left[R_{\text{emp}}\left(\hat{f}_i, S^i\right) - R[\hat{f}_i]\right]\right| \leq& \left|R_{\text{emp}}[\hat{f}, S] - R_{\text{emp}}\left(\hat{f}_i, S^i\right)\right| \\ &+ \left|R[\hat{f}] - R[\hat{f}_i]\right| \\ \leq& 4\beta + \frac{2M}{n}\end{aligned} \tag{33}$$

Thus, by the Theorem 3.3, we have $c_i = 4\beta + \frac{2M}{n}$, that,

$$\begin{aligned}\mathbb{P}\left\{|R_{\text{emp}}[\hat{f}, S] - R[\hat{f}] - 2\beta| > \varepsilon\right\} \leq & \mathbb{P}\left\{|R_{\text{emp}}[\hat{f}, S] - R[\hat{f}]| > \varepsilon + 2\beta\right\} \\ \leq & 2\exp\left(-\frac{n\varepsilon^2}{2(2n\beta + M)^2}\right)\end{aligned} \tag{34}$$

□

Within the above two theorems, we can directly get the following practical consequence.

Corollary (Uniform Convergence Bounds for RKHS)**.** *The algorithm minimizing the regularized risk $R_{reg}[f]$, as in (6), and with the assumptions of Theorem 3.2 and 3.4, we obtain*

$$\mathbb{P}\left\{\left|R_{emp}[\hat{f}] - R[\hat{f}]\right| > \varepsilon + 2\beta\right\} \leq 2\exp\left(-\frac{n}{2}\left(\frac{\varepsilon}{M}\right)^2\left(1 + \frac{4}{\lambda M}(C\kappa)^2\right)^{-2}\right) \tag{35}$$

where

$$\beta = \frac{2C^2\kappa^2}{n\lambda}$$

Remark. For practical considerations, (35) may be very useful, even if the rates are not optimal, since the bound is predictive even for small sample sizes and moderate regularization strength. Still, we expect that the constants

Comments

The idea of the discussion content was inspired by (Hofmann et al., 2008, Section 3) review, and the report is organized follow (Schölkopf and Smola, 2001, Chapter 12) structure, and so does the main proof ideas.

References

Bousquet, O. and Elisseeff, A. (2001). Algorithmic Stability and Generalization Performance. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press.

Bousquet, O. and Elisseeff, A. (2002). Stability and Generalization. *Journal of Machine Learning Research*, 2(Mar):499–526.

Hofmann, T., Schölkopf, B., and Smola, A. J. (2008). Kernel Methods in Machine Learning. *The Annals of Statistics*, 36(3):1171–1220.

McDiarmid, C. (1989). On the Method of Bounded Differences. In Siemons, J., editor, *Surveys in Combinatorics, 1989: Invited Papers at the Twelfth British Combinatorial Conference*, London Mathematical Society Lecture Note Series, pages 148–188. Cambridge University Press, Cambridge.

Ofer Dekel and Thach Nguyen (2011). Algorithmic Stability.

Schölkopf, B. and Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning series. MIT Press, Cambridge, MA, USA.

Stéphane Canu (2014). Lecture 4: Kernels and Associated Functions.