Note on the generalization bounds of the empirical risk minimizer

Satoshi Hayakawa

March 12, 2019

Abstract

We give a generalization error bound as a modification of Lemma 10 in Schmidt-Hieber (2017)[*1].

1 Settings and notations

Let us consider the following regression model. We observe i.i.d. random variables (X_i, Y_i) generated by

$$Y_i = f^\circ(X_i) + \xi_i, \qquad i = 1, 2, \ldots, n. \tag{1}$$

Here each ξ_i is a sampling noise independent of other variables. In this paper, we use the settings such that

- each X_i is d-dimensional and uniformly distributed in $[0,1]^d$,
- each Y_i is 1-dimensional,
- and ξ_i's are i.i.d. centerd Gaussian with variance σ^2 ($\sigma > 0$).

For simplicity, we sometimes use the notation $X^n := (X_1, \ldots, X_n)$, $Y^n := (Y_1, \ldots, Y_n)$ and $Z^n := (X_i, Y_i)_{i=1}^n$.

Definition 1.1. An *estimator* taking values in $\mathcal{F} \subset L^2([0,1]^d)$ is a measurable (with respect to Borel σ-algebra) map

$$(\mathbb{R}^d \times \mathbb{R})^n \to \mathcal{F}, \quad (X_i, Y_i)_{i=1}^n \mapsto \widehat{f}.$$

Remark 1.2. In the following, we often write only $\widehat{f}$ where we should write $(X_i, Y_i)^n \mapsto \widehat{f}$. For example, $\inf_{(X_i,Y_i)_{i=1}^n \mapsto \widehat{f} \in \mathcal{F}}$ is just denoted by $\inf_{\widehat{f} \in \mathcal{F}}$. Also, in the case $\mathcal{F} = L^2([0,1]^d)$, we omit $\mathcal{F}$, such as $\inf_{\widehat{f}}$.

To evaluate the quality of estimators, we have to adopt some indicator. For a fixed f° and a function $f \in L^2([0,1]^d)$, we have

$$\begin{aligned}
\mathrm{E}[(f(X_i) - Y_i)^2] &= \mathrm{E}[(f(X_i) - f^\circ(X_i))^2] - 2\mathrm{E}[\xi_i(f(X_i) - f^\circ(X_i))] + \mathrm{E}[\xi_i^2] \\
&= \mathrm{E}[(f(X_i) - f^\circ(X_i))^2] + \sigma^2 \\
&= \|f - f^\circ\|_{L^2}^2 + \sigma^2.
\end{aligned}$$

It means that how small the expected error $\mathrm{E}[(f(X_i) - Y_i)^2]$ is depends only on how small the L^2 distance $\|f - f^\circ\|_{L^2}^2$ is. It leads to the following definition of an indicator.

Definition 1.3. The L^2 *risk* for an estimator $\widehat{f}$ when is defined as

$$R(\widehat{f}, f^\circ) := \mathrm{E}\left[\|\widehat{f} - f^\circ\|_{L^2}^2\right].$$

We evaluate the quality of an estimator $\widehat{f}$ by this L^2 risk.

[*1] In the latest version of the paper, the technical flaw has been already fixed.

Remark 1.4. We omit n from notations because it is treated as a constant when we consider one regression problem.

To evaluate the convergence rate of an estimator, some "complexity" measure of the model is required. Here, we employ the ε-entropy for such a complexity measure.

Definition 1.5. (van der Vaart & Wellner 1996, Yang & Barron 1999) For a metric space (S, d) and $\varepsilon > 0$, a finite set $U \subset \overline{S}$ is called *ε-covering* if for any $x \in S$ there exists $y \in U$ such that $d(x, y) \leq \varepsilon$, and the logarithm of the minimum cardinality of ε-covering is called *covering ε-entropy* and denoted by $V_{(S,d)}(\varepsilon)$. Here, $\overline{S}$ is the completion of S with respect to the metric d.

2 Generalization bounds

The following theorem is useful for evaluating the convergence rate of the empirical risk minimizer.

Theorem 2.1. (Schmidt-Hieber 2017, Lemma 11) *In the Gaussian regression model (1), let $\widehat{f}$ be the empirical risk minimizer taking values in $\mathcal{F} \subset L^2([0,1]^d)$. Suppose every element $f \in \mathcal{F}$ satisfies $\|f\|_{L^\infty} \leq F$ for some fixed $F > 0$. Then, for an arbitrary $\delta > 0$, if $V_{(\mathcal{F},\|\cdot\|_{L^\infty})}(\delta) \geq 1$, then*

$$R(\widehat{f}, f^\circ) \leq 4 \inf_{f \in \mathcal{F}} \|f - f^\circ\|_{L^2}^2 + C\left(\frac{(F^2 + \sigma^2)V_{(\mathcal{F},\|\cdot\|_{L^\infty})}(\delta)}{n} + (F + \sigma)\delta\right)$$

holds, where $C > 0$ is an absolute constant.

Proof. (mainly following the original proof *2) First, we evaluate the value of

$$D := \left|\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n (\widehat{f}(X_i) - f^\circ(X_i))^2\right] - R(\widehat{f}, f^\circ)\right|.$$

Let $X_1', \ldots, X_n'$ be i.i.d. random variables generated to be independent from $(X_i, Y_i)_{i=1}^n$. Then we have

$$R(\widehat{f}, f^\circ) = \frac{1}{n}\sum_{i=1}^n \mathrm{E}\left[(\widehat{f}(X_i') - f^\circ(X_i'))^2\right],$$

so that we get

$$\begin{aligned} D &= \left|\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n \left((\widehat{f}(X_i) - f^\circ(X_i))^2 - (\widehat{f}(X_i') - f^\circ(X_i'))^2\right)\right]\right| \\ &\leq \frac{1}{n}\mathrm{E}\left[\left|\sum_{i=1}^n \left((\widehat{f}(X_i) - f^\circ(X_i))^2 - (\widehat{f}(X_i') - f^\circ(X_i'))^2\right)\right|\right]. \end{aligned}$$

Here, let $G_\delta = \{f_1, \ldots, f_N\}$ be a δ-covering of $\mathcal{F}$ with the minimum cardinality in L^∞ metric. Notice $\log N \geq 1$. If we define $g_j(x, x') := (f_j(x) - f^\circ(x))^2 - (f_j(x') - f^\circ(x'))^2$ and a random variable J taking values in $\{1, \ldots, N\}$ such that $\|\widehat{f} - f_J\|_{L^\infty} \leq \delta$, we have

$$D \leq \frac{1}{n}\mathrm{E}\left[\left|\sum_{i=1}^n g_J(X_i, X_i')\right|\right] + 8F\delta. \tag{2}$$

In the abouve evaluation, we have used the inequality

$$\left|(\widehat{f}(x) - f^\circ(x))^2 - (f_J(x) - f^\circ(x))^2\right| = \left|\widehat{f}(x) - f_J(x)\right|\left|\widehat{f}(x) + f_J(x) - 2f^\circ(x)\right| \leq 4F\delta.$$

*2 We noticed and fixed some technical flaws in the original proof.

Define constants $r_j := \max\{A, \|f_j - f^\circ\|_{L^2}\}$ $(j = 1, \ldots, N)$ and a random variable

$$T := \max_{1\le j\le N} \left| \sum_{i=1}^n \frac{g_j(X_i, X_i')}{r_j} \right|,$$

where $A > 0$ is a deterministic quantity fixed afterward. Then we have, since (2)

$$D \le \frac{1}{n}\mathrm{E}[r_J T] + 8F\delta \le \frac{1}{n}\sqrt{\mathrm{E}[r_J^2]\mathrm{E}[T^2]} + 8F\delta \le \frac{1}{2}\mathrm{E}[r_J^2] + \frac{1}{2n^2}\mathrm{E}[T^2] + 8F\delta \tag{3}$$

by Cauchy-Schwarz inequality and AM-GM inequality. Here, by the definition of J, $\mathrm{E}[r_J^2]$ can be evaluated as follows:

$$\begin{aligned}\mathrm{E}[r_J^2] \le A^2 + \mathrm{E}\left[\|f_J - f^\circ\|_{L^2}^2\right] &\le A^2 + \mathrm{E}\left[\|\widehat{f} - f^\circ\|_{L^2}^2\right] + 4F\delta \\ &= R(\widehat{f}, f^\circ) + A^2 + 4F\delta. \end{aligned}\tag{4}$$

Because of the independence of defined random variables,

$$\begin{aligned}\mathrm{E}\left[\left(\sum_{i=1}^n \frac{g_j(X_i, X_i')}{r_j}\right)^2\right] &= \sum_{i=1}^n \mathrm{E}\left[\left(\frac{g_j(X_i, X_i')}{r_j}\right)^2\right] \\ &= \sum_{i=1}^n \left(\mathrm{E}\left[\frac{(f_j(X_i) - f^\circ(X_i))^4}{r_j^2}\right] + \mathrm{E}\left[\frac{(f_j(X_i') - f^\circ(X_i'))^4}{r_j^2}\right]\right) \\ &\le 2F^2 n\end{aligned}$$

holds, where we have used the fact that each $g_j(X_i, X_i')$ is centered. Then, using Bernstein's inequality (Theorem **??**), we have, in terms of $r := \min_{1\le j\le N} r_j$,

$$\mathrm{P}(T^2 \ge t) = \mathrm{P}(T \ge \sqrt{t}) \le 2N \exp\left(-\frac{t}{2F^2\left(2n + \frac{\sqrt{t}}{3r}\right)}\right), \quad t \ge 0.$$

Let us evaluate $\mathrm{E}[T^2]$. For arbitrary $t_0 > 0$, it holds that

$$\begin{aligned}\mathrm{E}[T^2] = \int_0^\infty P(T^2 \ge t)\,\mathrm{d}t &\le t_0 + \int_{t_0}^\infty P(T^2 \ge t)\,\mathrm{d}t \\ &\le t_0 + 2N\int_{t_0}^\infty \exp\left(-\frac{t}{8F^2 n}\right)\,\mathrm{d}t + 2N\int_{t_0}^\infty \exp\left(-\frac{3r\sqrt{t}}{4F^2}\right)\,\mathrm{d}t.\end{aligned}$$

We compute these two integration values in terms of t_0 :

$$\begin{aligned}\int_{t_0}^\infty \exp\left(-\frac{t}{8F^2 n}\right)\,\mathrm{d}t &= \left[-8F^2 n \exp\left(-\frac{t}{8F^2 n}\right)\right]_{t_0}^\infty = 8F^2 n \exp\left(-\frac{t_0}{8F^2 n}\right), \\ \int_{t_0}^\infty \exp\left(-\frac{3r\sqrt{t}}{4F^2}\right)\,\mathrm{d}t &= \int_{t_0}^\infty \exp(-a\sqrt{t})\,\mathrm{d}t \qquad (a := 3r/4F^2) \\ &= \left[-\frac{2(a\sqrt{t}+1)}{a^2}\exp(-a\sqrt{t})\right]_{t_0}^\infty \\ &= \frac{8F^2\sqrt{t_0}}{3r}\exp\left(-\frac{3r\sqrt{t_0}}{4F^2}\right) + \frac{32F^2}{9r^2}\exp\left(-\frac{3r\sqrt{t_0}}{4F^2}\right).\end{aligned}$$

Now we determine $A = \sqrt{t_0}/6n$. Since we have $r \geq A = \sqrt{t_0}/6n$,

$$\begin{aligned}\mathrm{E}[T^2] &\leq t_0 + 2N\left(8F^2n + 16F^2n + \frac{128F^2n^2}{t_0}\right)\exp\left(-\frac{t_0}{8F^2n}\right)\\ &\leq t_0 + 16NF^2n\left(3 + \frac{16n}{t_0}\right)\exp\left(-\frac{t_0}{8F^2n}\right)\end{aligned}$$

holds. Let $t_0 = 8F^2n\log N$, then the above evaluation is rewritten as

$$\mathrm{E}[T^2] \leq 8F^2n\left(\log N + 6 + \frac{2}{F^2\log N}\right). \tag{5}$$

Finally, we combine (3), (4), (5) and $A^2 = \dfrac{2F^2\log N}{9n}$ to get

$$\begin{aligned}D &\leq \left(\frac{1}{2}R(\widehat{f}, f^\circ) + \frac{1}{2}A^2 + 2F\delta\right) + \frac{4F^2}{n}\left(\log N + 6 + \frac{2}{F^2\log N}\right) + 8F\delta\\ &\leq \frac{1}{2}R(\widehat{f}, f^\circ) + \frac{F^2}{n}\left(\frac{37}{9}\log N + 32\right) + 10F\delta,\end{aligned}$$

where we have used the fact that $\log N \geq 1$. So we get an evaluation

$$R(\widehat{f}, f^\circ) \leq 2\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\widehat{f}(X_i) - f^\circ(X_i))^2\right] + \frac{2F^2}{n}\left(\frac{37}{9}\log N + 32\right) + 20F\delta. \tag{6}$$

Next we evaluate the quantity

$$\widehat{R} := \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\widehat{f}(X_i) - f^\circ(X_i))^2\right]. \tag{7}$$

Since $\widehat{f}$ is an empirical risk minimizer, for arbitrary $f \in \mathcal{F}$,

$$\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\widehat{f}(X_i) - Y_i)^2\right] \leq \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2\right]$$

holds. As $Y_i = f^\circ(X_i) + \xi_i$, we have

$$\begin{aligned}&\mathrm{E}\left[(f(X_i) - Y_i)^2\right] - \mathrm{E}\left[(\widehat{f}(X_i) - Y_i)^2\right]\\ &= \mathrm{E}\left[(f(X_i) - f^\circ(X_i))^2\right] - 2\mathrm{E}\left[\xi_i f(X_i)\right] - \mathrm{E}\left[(\widehat{f}(X_i) - f^\circ(X_i))^2\right] + 2\mathrm{E}\left[\xi_i\widehat{f}(X_i)\right]\\ &= \left(\|f - f^\circ\|_{L^2}^2 + 2\mathrm{E}\left[\xi_i\widehat{f}(X_i)\right]\right) - \mathrm{E}\left[(\widehat{f}(X_i) - f^\circ(X_i))^2\right].\end{aligned}$$

Here we have used the fact that

$$\mathrm{E}[\xi_i f(X_i)] = \mathrm{E}[\xi_i]\mathrm{E}[f(X_i)] = 0$$

holds because of the independence between ξ_i and X_i, and the fact that both ξ_i and $f(X_i)$ have a finite L^1 norm. So we have

$$\widehat{R} \leq \|f - f^\circ\|_{L^2}^2 + \mathrm{E}\left[\frac{2}{n}\sum_{i=1}^{n}\xi_i\widehat{f}(X_i)\right]. \tag{8}$$

Let us evaluate the second term in RHS.

$$\begin{aligned}\left|\mathrm{E}\left[\frac{2}{n}\sum_{i=1}^n \xi_i \widehat{f}(X_i)\right]\right| &= \left|\mathrm{E}\left[\frac{2}{n}\sum_{i=1}^n \xi_i(\widehat{f}(X_i) - f^\circ(X_i))\right]\right| \\ &\leq \frac{2\delta}{n}\mathrm{E}\left[\sum_{i=1}^n |\xi_i|\right] + \left|\mathrm{E}\left[\frac{2}{n}\sum_{i=1}^n \xi_i(f_J(X_i) - f^\circ(X_i))\right]\right|. \end{aligned} \tag{9}$$

Here, the first term is upper bounded by using Cauchy-Schwarz inequality:

$$\frac{2\delta}{n}\mathrm{E}\left[\sum_{i=1}^n |\xi_i|\right] \leq \frac{2\delta}{n}\mathrm{E}\left[n^{1/2}\left(\sum_{i=1}^n \xi_i^2\right)^{1/2}\right] \leq \frac{2\delta}{\sqrt{n}}\mathrm{E}\left[\sum_{i=1}^n \xi_i^2\right]^{1/2} = 2\sigma\delta. \tag{10}$$

Let ε_j $(j = 1, \dots, N)$ be random variables defined as

$$\varepsilon_j := \frac{\sum_{i=1}^n \xi_i(f_j(X_i) - f^\circ(X_i))}{\left(\sum_{i=1}^n (f_j(X_i) - f^\circ(X_i))^2\right)^{1/2}},$$

where $\varepsilon_j := 0$ if the denominator equals to 0. Notice each ε_j follows a centered Gaussian with variance σ^2 (conditionally on $X_1, \dots, X_n$). Now we have, using Cauchy-Schwarz inequality and AM-GM inequality,

$$\begin{aligned}\left|\mathrm{E}\left[\frac{2}{n}\sum_{i=1}^n \xi_i(f_J(X_i) - f^\circ(X_i))\right]\right| &= \frac{2}{n}\left|\mathrm{E}\left[\left(\sum_{i=1}^n (f_J(X_i) - f^\circ(X_i))^2\right)^{1/2}\varepsilon_J\right]\right| \\ &\leq \frac{2}{\sqrt{n}}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n (f_J(X_i) - f^\circ(X_i))^2\right]^{1/2}\mathrm{E}\left[\max_{1\leq j\leq N}\varepsilon_j^2\right]^{1/2} \\ &\leq \frac{2}{\sqrt{n}}\sqrt{\widehat{R} + 4F\delta}\ \mathrm{E}\left[\max_{1\leq j\leq N}\varepsilon_j^2\right]^{1/2} \\ &\leq \frac{1}{2}(\widehat{R} + 4F\delta) + \frac{2}{n}\mathrm{E}\left[\max_{1\leq j\leq N}\varepsilon_j^2\right]. \end{aligned} \tag{11}$$

By a similar argument as in the proof of Lafferty, Liu & Wasserman (2008, Theorem 7.47), for any $0 < t < 1/2\sigma^2$,

$$\begin{aligned}\exp\left(t\mathrm{E}\left[\max_{1\leq j\leq N}\varepsilon_j^2\right]\right) &\leq \mathrm{E}\left[\max_{1\leq j\leq N}\exp\left(t\varepsilon_j^2\right)\right] \qquad \text{(by Jensen's inequality)} \\ &\leq N\mathrm{E}\left[\exp\left(t\varepsilon_1^2\right)\right] \\ &= \frac{N}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{tx^2}e^{-\frac{x^2}{2\sigma^2}}\,\mathrm{d}x = \frac{N}{\sqrt{1-2\sigma^2 t}}\end{aligned}$$

holds. So we have, by determining $t = 1/4\sigma^2$,

$$\mathrm{E}\left[\max_{1\leq j\leq N}\varepsilon_j^2\right] \leq 4\sigma^2\log(\sqrt{2}N) \leq 4\sigma^2(\log N + 1) \tag{12}$$

Now we combine (8)–(12) to get

$$\widehat{R} \leq \|f - f^\circ\|_{L^2}^2 + 2\sigma\delta + \frac{1}{2}(\widehat{R} + 4F\delta) + \frac{8\sigma^2}{n}(\log N + 1),$$

so that

$$\widehat{R} \leq 2\|f - f^\circ\|_{L^2}^2 + 4(\sigma + F)\delta + \frac{16\sigma^2}{n}(\log N + 1) \tag{13}$$

holds.

Finally, since f is an arbitrary element of $\mathcal{F}$, we combine (6), (7) and (13) to have

$$R(\widehat{f}, f^\circ) \leq 4 \inf_{f \in \mathcal{F}} \|f - f^\circ\|_{L^2}^2 + \frac{1}{n}\left(\left(\frac{37}{9}F^2 + 32\sigma^2\right)\log N + 32(F^2 + \sigma^2)\right)$$
$$+ (18F + 8\sigma)\delta,$$

and this leads to the conclusion. $\square$

References

Lafferty, J., Liu, H. & Wasserman, L. (2008). Concentration on measure, `http://www.stat.cmu.edu/~larry/=sml/Concentration.pdf`.

Schmidt-Hieber, J. (2017). Nonparametric regression using deep neural networks with ReLU activation function, *arXiv preprint arXiv:1708.06633 (ver.3)* .

van der Vaart, A. W. & Wellner, J. A. (1996). *Weak convergence and empirical processes*, Springer.

Yang, Y. & Barron, A. (1999). Information-theoretic determination of minimax rates of convergence, *The Annals of Statistics* **27**(5): 1564–1599.