Summary of Theory

Miaoyan Wang, Oct 4, 2020

1 Set-up

Consider the linear function class

$$\mathcal{F}(d,r,s) = \{\boldsymbol{X} \mapsto \langle \boldsymbol{X}, \boldsymbol{B}\rangle \bigm| \text{rank}(\boldsymbol{B}) \leq r,\ \text{Supp}(\boldsymbol{B}) \leq s,\ \boldsymbol{B} \in \mathbb{R}^{d\times d}\}. \tag{1}$$

For any function $f \in \mathcal{F}(d,r,s)$, we define $\|f\|_F = \|\boldsymbol{B}\|_F$. (do I need to care about $\boldsymbol{X}$?? consistent with RHKS?). Distinguish to L-2 norm of f?

Let $\{(\boldsymbol{X}_i, y_i) \in \mathbb{R}^{d\times d} \times \{\pm 1\} \colon i = 1,\ldots,n\}$ denote the i.i.d. training sample from an unknown distribution $\mathbb{P}(\boldsymbol{X}, y)$. We are interested in the high-dimensional regime as $n, d \to \infty$, while holding s, r as fixed constants.

Define the restricted eigenvalue as

$$\lambda_{\max}(\boldsymbol{X}) = \max_{\boldsymbol{x}\in \boldsymbol{S}^d, \|\boldsymbol{x}\|_0 \leq s} \boldsymbol{a}^T \boldsymbol{X} \boldsymbol{a}.$$

Assume there exists a constant $C > 0$ such that $\lambda_{\max}(\boldsymbol{X}) \leq C$, a.s. as $d \to \infty$. The restricted eigenvalue condition incorporates (i) bounded feature $\|\boldsymbol{X}\|_F \leq C$, and (ii) Gaussian feature $\boldsymbol{X}$ with i.i.d. $N(0,1)$ entries. In the later case, the feature space is unbounded, $\|\boldsymbol{X}\|_F \asymp \mathcal{O}(d)$ as $d \to \infty$, but spectral norm is bounded $\lambda_{\max}(\boldsymbol{X}) \asymp s = \mathcal{O}(1)$. We also need $\|\boldsymbol{B}\|_F \leq \max_{\boldsymbol{X}\in\mathbb{R}^{d\times d}} \langle \boldsymbol{B}, \boldsymbol{X}\rangle$ (this is natural) do we need assumptions on $\boldsymbol{X}$ in the probability estimation?? We need $\|f\|_\infty$ bounded in the L_2 entropy. We require $\boldsymbol{X}$ spread out, in particular, cover the set of $\boldsymbol{B}$

2 Theory

Definition 1 (Classification risk and surrogate risk)**.** Let $f(\cdot)\colon \mathbb{R}^{d\times d} \mapsto \mathbb{R}$ be the decision function of interest, $\ell(\cdot)\colon \mathbb{R} \mapsto \mathbb{R}_{\geq 0}$ be a surrogate loss function in terms of the margin $yf(\boldsymbol{X})$. We define the 0/1 classification risk and surrogate risk,

$$R(f) = \mathbb{P}(y \neq \text{sign} f(\boldsymbol{X})), \quad R_\ell(f) = \mathbb{E}\ell(yf(\boldsymbol{X})).$$

Assumption 1 (Surrogate loss)**.** Assume that the surrogate loss ℓ satisfies the following two conditions

i. ℓ is a L-Lipschitz function and ℓ entrywise dominates the 0/1 loss. This assumption implies that $R(f) \leq R_\ell(f)$ for all functions f.

ii. The loss is Fisher consistent,

$$R(f_{\text{bayes}}) = R_\ell(f_{\text{bayes}}), \quad \text{and} \quad \mathop{\arg\min}_{\text{all possible } f} R_\ell(f) = \mathop{\arg\min}_{\text{all possible } f} R(f). \tag{2}$$

That is, replacing 0/1 loss by surrogate loss does not change the minimal risk and minimizer. Note that the right hand side of (2) is obtained at $f_{\text{bayes}}(\cdot)\colon \boldsymbol{X} \mapsto \text{sign}\{\mathbb{P}(y=1|\boldsymbol{X}) - 1/2\}$ a.s. except for the decision boundary $\{\boldsymbol{X}\colon \mathbb{P}(y=1|\boldsymbol{X}) = 1/2\}$. (where do we use the global minimum assumption?)

We denote the empirical risks calculated from the training sample,

$$\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n}(y_i \neq \text{sign} f(\boldsymbol{X}_i)), \quad \hat{R}_\ell(f) = \frac{1}{n}\sum_{i=1}^{n}\ell(y_i f(\boldsymbol{X}_i)).$$

Proposition 1 (Generalization)**.** Consider the function class $\mathcal{F}(d,r,s)$ in (1) and a surrogate loss ℓ under Assumption 1i. With very high probability over $\{(\boldsymbol{X}_i, y_i)\}$, we have

$$\sup_{f\in\mathcal{F}(d,r,s)} |R_\ell(f) - \hat{R}_\ell(f)| \leq CLs\log d\sqrt{\frac{r}{n}} \sup_{\boldsymbol{B}\in\mathcal{F}(d,r,s)} \|\boldsymbol{B}\|_F.$$

Corollary 2 (Excess risk)**.** Consider the same assumptions as in Proposition 1. Let $\hat{f} \in \mathcal{F}(d,r,s)$ denote empirical surrogate risk minimizer constrained to the function class $\mathcal{F}(d,r,s)$,

$$\hat{f} = \mathop{\arg\min}_{f\in\mathcal{F}(d,r,s)} \frac{1}{n}\sum_{i=1}^{n}\ell(y_i f(\boldsymbol{X}_i)).$$

Then with very high probability over the training set, the surrogate risk error satisfies

$$R_\ell(\hat{f}) - \inf_{f\in\mathcal{F}(d,r,s)} R_\ell(f) \leq 2CLs\log d\sqrt{\frac{r}{n}} \sup_{\boldsymbol{B}\in\mathcal{F}(d,r,s)} \|\boldsymbol{B}\|_F.$$

and the classification error satisfies

$$R(\hat{f}) - \inf_{f\in\mathcal{F}(d,r,s)} R(f) \leq 2CLs\log d\sqrt{\frac{r}{n}} \sup_{\boldsymbol{B}\in\mathcal{F}(d,r,s)} \|\boldsymbol{B}\|_F.$$

Write in terms of Bayes error?

Remark 1 (Two shortcomings)**.** First, the convergence in sample size is of order $\mathcal{O}(1/\sqrt{n})$. This can be improved to $\mathcal{O}(1/n)$ upon mean-variance (low noise) conditions. Second, the generalization bound depends on $\sup_{\boldsymbol{B}\in\mathcal{F}(d,r,s)}\|\boldsymbol{B}\|_F$, which may depends on d. In fact, we can use an adaptive penalization to replace this term to $\|\boldsymbol{B}^*\|_F$, where $\boldsymbol{B}^*$ is the matrix that induces f^*; i.e., $f^*(\boldsymbol{X}) = \langle \boldsymbol{X}, \boldsymbol{B}^*\rangle$.

To overcome the above pitfalls, we propose to use the following penalized empirical surrogate risk minimizer

$$\hat{f}_\lambda = \operatorname*{arg\,min}_{f\in\mathcal{F}(d,r,s)} \left\{ \frac{1}{n}\sum_{i=1}^{n} \ell(y_i f(\boldsymbol{X}_i)) + \lambda \|f\|_F^2 \right\}, \tag{3}$$
$$\text{where}\quad \mathcal{F}(d,r,s) = \{\boldsymbol{X} \mapsto \langle \boldsymbol{X}, \boldsymbol{B}\rangle \mid \text{rank}(\boldsymbol{B}) \le r,\ \text{Supp}(\boldsymbol{B}) \le s,\ \boldsymbol{B}\in\mathbb{R}^{d\times d}\}.$$

Assumption 2. Consider the following assumptions:

i. As $n, d \to \infty$, there exists a sequence $f_n^* \in \mathcal{F}(d,r,s)$ that is (1) bounded $\|f_n^*\|_F^2 \le J^*$ for some constant $C > 0$, and (2) $R_\ell(f_n^*) - R_\ell(f_{\text{bayes}}) \le a_n \to 0$. The upper bound J^* is allowed to depend on $d = d(n)$.

ii. There exist constants $a > 0$ and $\rho \in [0,1]$, such that, for any sufficient small $\delta > 0$,

$$\text{Var}\left[\ell(yf(\boldsymbol{X})) - \ell(yf_{\text{bayes}}(\boldsymbol{X}))\right] \le a\mathbb{E}\left[\ell(yf(\boldsymbol{X})) - \ell(yf_{\text{bayes}}(\boldsymbol{X}))\right]^\rho$$

holds for all $f \in \{f\in\mathcal{F}(d,r,s) \colon R_\ell(f) - R_\ell(f_{\text{bayes}}) \le \delta\}$ in a δ-neighborhood of f_{bayes}.

iii. There exist constants $b > 0$ and $\alpha \ge 0$, such that, for any sufficient small $\delta > 0$,

$$\mathbb{E}\big|\text{sign} f(\boldsymbol{X}) - \text{sign} f_{\text{bayes}}(\boldsymbol{X})\big| \le b\left[R_\ell(f) - R_\ell(f_{\text{bayes}})\right]^\alpha$$

holds for all $f \in \{f\in\mathcal{F}(d,r,s) \colon R_\ell(f) - R_\ell(f_{\text{bayes}}) \le \delta\}$ in a δ-neighborhood of f_{bayes}.

Theorem 3 (Main result for classification accuracy)**.** Suppose Assumption 1 and Assumption 2i-2ii hold. Consider a penalized empirical surrogate risk minimizer (3) with the regularity parameter $\lambda \asymp \frac{1}{J^*}\left(\frac{rs\log d}{n}\right)^{1/(2-\rho)}$. We have that, with very high probability, the surrogate error and classification error satisfy

$$R_\ell(\hat{f}_\lambda) - R_\ell(f_{\text{bayes}}) \le C\left(\frac{rs\log d}{n} + a_n\right)^{1/(2-\rho)}, \quad R(\hat{f}_\lambda) - R(f_{\text{bayes}}) \le C\left(\frac{rs\log d}{n} + a_n\right)^{1/(2-\rho)}. \tag{4}$$

Furthermore, suppose Assumption 2iii holds. Then, Assumption 2ii holds with $\rho = \alpha \wedge 1$, and the estimation error for level sets satisfies

$$\mathbb{P}(\hat{\boldsymbol{X}}\Delta\boldsymbol{X}_{\text{bayes}}) \le C\left(\frac{rs\log d}{n} + a_n\right)^{\alpha/(2-\alpha\wedge 1)},$$

where $\hat{\boldsymbol{X}} = \{\boldsymbol{X} \colon \hat{f}(\boldsymbol{X}) \ge 0\}$ and $\boldsymbol{X}_{\text{bayes}} = \{\boldsymbol{X} \colon f_{\text{bayes}}(\boldsymbol{X}) \ge 0\}$ are estimated and true level sets, respectively.

Remark 2. The bound (4) implies the consistency of risk estimator in the ultra high-dimensional regime. In particular, the dimension of feature matrices d is allowed to grow sub-exponentially in sample size n; i.e., $d = o(e^n)$.

Theorem 4 (Main result for probability estimation)**.** Consider the same assumptions of Theorem 3. For the estimator $\hat{p}\colon \mathbb{R}^{d\times d} \to [0,1]$ obtained from level-set estimation,

$$\hat{p}(\boldsymbol{X}) = \frac{1}{H}\sum_{h=1}^{H} \mathbb{1}\{\boldsymbol{X}\colon \hat{f}_h(\boldsymbol{X}) \geq 0\}, \quad \text{for all } \boldsymbol{X} \in \mathbb{R}^{d\times d}.$$

With probability at least $1 - C\exp(-an\lambda^{2-\alpha})$,

$$\mathbb{E}|\hat{p}(\boldsymbol{X}) - p(\boldsymbol{X})| \geq \underbrace{\frac{1}{2H}}_{\text{discretization error}} + \frac{a(H+1)}{2}\mathcal{O}\left(\underbrace{\frac{rs\log d}{n}}_{\text{statistical error}} + \underbrace{a_n}_{\text{approximation error}}\right)^{\alpha^2}.$$

Corollary 5. Assume $a_n \leq \frac{rs\log d}{n}$. Choosing $H \asymp (\frac{rs\log d}{n})^{-\alpha^2/2}$ gives the estimation error,

$$\mathbb{E}|\hat{p}(\boldsymbol{X}) - p(\boldsymbol{X})| \leq C\left(\frac{rs\log d}{n}\right)^{\rho/2}.$$

3 Algorithm

4 Numerical experiments

Three goals

1. Assess the 0/1 and hinge loss errors
2. Assess the level-set estimation
3. Assess the probability estimation

5 Proofs

Proof of Proposition 1. It suffices to bound the Radamecher complexity

$$\begin{aligned}
\mathrm{Rad}(f) &= \frac{1}{n}\mathbb{E}\max_{f\in\mathcal{F}(d,r,s)}\sum_{i\in[n]}\sigma_i\langle\boldsymbol{X}_i,\boldsymbol{B}\rangle \\
&\leq \frac{1}{n}\|\boldsymbol{B}\|_*\|\sum_{i\in[n]}\sigma_i\boldsymbol{X}_i\|_{\text{sp}} \\
&\leq \frac{1}{n}\mathbb{E}\max_{\text{possible choice } \mathcal{S} \text{ from } [d]}\ \max_{\text{given set } \mathcal{S}}\langle\boldsymbol{B},\sum_i\sigma_i\boldsymbol{X}_i\rangle \\
&\leq \frac{1}{n}\log\binom{d}{s}\mathbb{E}\max_{\{\boldsymbol{a}_r,\boldsymbol{b}_r,\lambda_r\}}\langle\sum_r\lambda_r\boldsymbol{a}_r\boldsymbol{b}_r^T,\ \sum_i\sigma_i\boldsymbol{X}_i\rangle
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{s}{n}\log d \sum_r \lambda_r \langle \boldsymbol{a}_r \boldsymbol{b}_r,\ \sum_i \sigma_i \boldsymbol{X}_i\rangle \\
&\leq \frac{s}{n}\log d \sqrt{n}\lambda_{\max}(\boldsymbol{X})\sqrt{r}\|\boldsymbol{B}\|_F \\
&\leq Cs\sqrt{\frac{r}{n}}\log d \|\boldsymbol{B}\|_F.
\end{aligned}
$$

□

Proof of Corollary 2. Consider the decomposition

$$
\begin{aligned}
R_\ell(\hat{f}) - \inf_{f\in\mathcal{F}(d,r,s)} R_\ell(f) &\leq R_\ell(\hat{f}) - \hat{R}_\ell(\hat{f}) + \hat{R}_\ell(\hat{f}) - \hat{R}_\ell(f^*) + \hat{R}_\ell(f^*) - R_\ell(f^*) \\
&\leq |R_\ell(\hat{f}) - \hat{R}_\ell(\hat{f})| + \hat{R}_\ell(f^*) - R_\ell(f^*) \\
&\leq 2\mathrm{Rad}(f) \\
&\leq 2CLs\sqrt{\frac{r}{n}}\log d \max_{f\in\mathcal{F}(d,r,s)} \|\boldsymbol{B}\|_F
\end{aligned}
$$

□

Proof of Theorem 3. We set $\delta_n = \mathcal{O}(\frac{rs\log d}{n})^\rho$. Then

$$
\begin{aligned}
&\mathbb{P}\left\{R_\ell(\hat{f}) - R_\ell(f_{\text{bayes}}) \geq \delta\right\} \\
&\leq \mathbb{P}\left\{\sup_{\{f\in\mathcal{F}(d,r,s)\colon R_\ell(f) - R_\ell(f_{\text{bayes}})\geq\delta\}} \left(\hat{R}(f^*) + \lambda\|f^*\|_F^2 - \hat{R}(f) - \lambda\|f\|_F^2\right) \geq 0\right\} \\
&\leq 3.5\exp(-an\lambda^{2-\rho}),
\end{aligned}
$$

where the last line comes from Lemma 7, by taking $\lambda \asymp \delta_n$. (where a_n enters?)

The classification bound follows by noting that $R(\hat{f}) \leq R_\ell(\hat{f})$ and $R(f_{\text{bayes}}) = R_\ell(f_{\text{bayes}})$. □

Definition 2 (bracketing number, uniform entropy, and bounded functions)**.** Consider a function set $\mathcal{F}$, and let $\varepsilon > 0$. We call $\{(f_m^l, f_m^u)\}_{m=1}^M$ an L_2-metric, ε-bracketing function set of $\mathcal{F}$, if for every $f \in \mathcal{F}$, there exists an $m \in [M]$ such that

$$
f_m^l(\boldsymbol{X}) \leq f(\boldsymbol{X}) \leq f_m^u(\boldsymbol{X}), \quad \text{for all } \boldsymbol{X} \in \mathbb{R}^{d\times d},
$$

and

$$
\|f_m^l - f_m^u\|_2 \stackrel{\text{def}}{=} \sqrt{\mathbb{E}|f_m^l(\boldsymbol{X}) - f_m^u(\boldsymbol{X})|^2} \leq \varepsilon,\ \text{ for all } m = 1,\ldots,M.
$$

The bracketing number with L_2-metric, $\mathcal{H}_{[\ ]}(\varepsilon,\ \mathcal{F},\ \|\cdot\|_2)$, is defined as the logarithm of the smallest cardinality of the ε-bracketing function set of $\mathcal{F}$. Furthermore, consider the set of functions with

L_∞ bound no larger than M, denoted $\mathcal{F}(M) = \{f \in \mathcal{F}\colon \|f\|_\infty \le M\}$. Then we have

$$\mathcal{H}_{[\ ]}(\varepsilon,\ \mathcal{F}(M),\ \|\cdot\|_2) \le \mathcal{H}(\varepsilon,\ \mathcal{F}(M),\ \|\cdot\|_\infty).$$

Lemma 6 (Uniform entropy for bounded functions in $\mathcal{F}(d,r,s)$)**.** Let $\mathcal{F}(d,r,s)$ denote the function class in (1). Consider the subset of functions with L_∞ bound no larger than M, denoted $\mathcal{F}(M) = \{f \in \mathcal{F}(d,r,s)\colon \|f\|_\infty \le M\}$ for $M = 1, 2, \ldots$. When ε sufficiently small, we have

$$\mathcal{H}(\varepsilon,\ \mathcal{F}(M),\ \|\cdot\|_\infty) \le 5rs \log \frac{M\sqrt{r}\lambda_{\max} d}{\varepsilon}.$$

Proof. For a given matrix $\boldsymbol{B}$, the definition $f(\boldsymbol{X}) = \langle \boldsymbol{X}, \boldsymbol{B}\rangle$ implies that $\|\boldsymbol{B}\|_F \le \max_{\boldsymbol{X}\in\mathbb{R}^{d\times d}} \langle \boldsymbol{X}, \boldsymbol{B}\rangle = \|f\|_\infty \le \|\boldsymbol{B}\|_F \sqrt{r}\lambda_{\max}$. Therefore, we have

$$\mathcal{H}(\varepsilon\sqrt{r}\lambda_{\max},\ \mathcal{F}(M),\ \|\cdot\|_\infty) = \mathcal{H}\left(\varepsilon,\ \mathcal{B}(M),\ \|\cdot\|_F\right), \tag{5}$$

where we have defined the matrix set $\mathcal{B}(M) = \{\boldsymbol{B} \in \mathbb{R}^{d\times d}\colon \mathrm{rank}(\boldsymbol{B}) \le r,\ \mathrm{Supp}(\boldsymbol{B}) \le s,\ \|\boldsymbol{B}\|_F \le M\}$. Note that the F-norm of matrix $\boldsymbol{B}$ is equivalent to the l_2-norm of the vector $\mathrm{vec}(\boldsymbol{B})$, and the vector l_2-norm is lower bounded by vector l_∞-norm in Euclidean space. Therefore, it suffices to bound $\mathcal{H}(\varepsilon,\ \mathcal{B}(M),\ \|\cdot\|_\infty)$. Now fix a subset $S \subset [d]$ with $|S| = s$, and let $\mathcal{B}_S(M) \subset \mathcal{B}(M)$ denote the subset of matrices satisfying $\boldsymbol{B}(i,j) = 0$ whenever $(i,j) \notin S^2$. Based on ..., the ε-covering of $\mathcal{B}_S(M)$ has entropy

$$\mathcal{H}(\varepsilon,\ \mathcal{B}_S(M),\ \|\cdot\|_\infty) \le r(2s+1)\log\left(\frac{M}{\varepsilon}\right).$$

In view of $\mathcal{B}(M) \subset \bigcup_{S\subset[d],|S|=s} \mathcal{B}_S(M)$, an ε-covering set $\mathcal{B}(M)$ is then given by the union of ε-covering set of $\mathcal{B}_S(M)$. Using Stirling's bound, we derive that

$$\mathcal{H}(\varepsilon,\ \mathcal{B}(M),\ \|\cdot\|_\infty) \le 2s\log\frac{d}{s} + r(2s+1)\log\frac{M}{\varepsilon} \le 5rs\log\frac{Md}{\varepsilon}.$$

Substituting ε by $\varepsilon/\sqrt{r}\lambda_{\max}$ into (5) concludes the proof. □

Lemma 7 (Metric of local $\mathcal{F}(d,r,s)$)**.** Let $\delta > 0$ be the solution to the following inequality,

$$\max_{M\ge 2}\left\{\frac{1}{\delta+\lambda(M/2-1)}\int_{\delta+\lambda(M/2-1)}^{(\delta+\lambda(M/2-1))^{\rho/2}} \sqrt{\mathcal{H}(\varepsilon,\ \mathcal{F}(M),\ \|\cdot\|_\infty)}d\varepsilon\right\} \le n^{1/2}.$$

Then we have $\delta = \mathcal{O}\left(\frac{rs\log d}{n}\right)^{\rho}$ provided that $\lambda \le 4\delta$ (??).

Theorem 8. Let $\mathcal{F}$ be a class of functions. Let $T, V \in (0,\infty)$ denote the upper bound of functions in $\mathcal{F}$ in L_∞ and L_2 norms; that is, $\sup_{f\in\mathcal{F}} \|f\|_\infty \le T$ and $\sup_{f\in\mathcal{F}} \mathrm{Var}(f) \le v$. Let $E_n(f) = \frac{1}{n}\sum_{i=1}^n (f(Y_i)$ be the empirical process. Define x_n^* be the solution of the equation to the following equation

$$\frac{1}{x}\int_{x}^{\sqrt{V}} \sqrt{\mathcal{H}_{[\,]}(\varepsilon, \mathcal{F}, \|\cdot\|_2)}d\varepsilon = \sqrt{n}.$$

Suppose the $\sqrt{V} \leq T$, and

$$x_n^* \lesssim \frac{V}{T}, \quad \text{and} \quad \mathcal{H}_{[\,]}(\sqrt{V}, \mathcal{F}, \|\cdot\|_2) \lesssim \frac{nx_n^*}{T}.$$

Then we have

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} E_n(f) \geq \mathbb{E}f(Y) + x_n^*\right) \lesssim \exp\left(-cnx_n^*\right).$$

Remark 3. The function set $\mathcal{F}$, and bounds T, v are allowed to depend on n.

We view $\mathcal{Y}_\Omega = \{\bar{\mathcal{Y}}(\omega)\colon \omega \in \Omega\}$ as a collection of n i.i.d. random variables where the randomness is induced form both $\bar{\mathcal{Y}}$ and $\omega \sim \Pi$, and view the tensor $\mathcal{Z}$ as a function that maps $\bar{\mathcal{Y}}_\Omega$ to $L(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega)$.

Specifically, the data takes the form $\{y_i\colon i = 1, 2, \ldots, n\}$, where for each $i \in [n]$, y_i is i.i.d. sampled from all entries of $\mathcal{Y}$ based on $\omega \sim \Pi$. We denote y_i i.i.d. random variables where the randomness is induced from both Π and noise in the tensor model.

The loss function then takes the form

$$L(\mathcal{Z}, \mathcal{Y}_\Omega) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(y_i),$$

where $\ell_i(y_i) = |y_i||\text{sign}(z_i) - \text{sign}(y_i)|$. The collection of function $\{\ell_i\colon i \in [d_1] \times \cdots [d_K]\}$ is one-to-one $\mathcal{Z}$. For notational simplicity, we write $L(\mathcal{Z})$ in place of $L(\mathcal{Z}, \mathcal{Y}_\Omega)$. The relevant probability statements, such as $\mathbb{E}$ and Var, are taken with respect to $\mathcal{Y}(\omega)$.

Because $\bar{\Theta}$ is the global minimizer of $\text{Risk}(\cdot)$, and by definition, $L(\hat{Z}) \leq L(\bar{\Theta})$, we have the following inclusion of the event

$$\{\text{Risk}(\hat{\mathcal{Z}}) - \text{Risk}(\bar{\Theta}) \geq T_n\} \subset \{\sup_{\mathcal{Z}\in\mathcal{F}} \big(\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta}) + L(\bar{\Theta}) - L(\mathcal{Z})\big) \geq T_n\}.$$

Therefore,

$$\mathbb{P}\left(\text{Risk}(\hat{\mathcal{Z}}) - \text{Risk}(\bar{\Theta}) \geq T_n\right) \leq \mathbb{P}\left(\sup_{\mathcal{Z}\in\mathcal{F}} |L(\mathcal{Z}) - \text{Risk}(\mathcal{Z}) - L(\bar{\Theta}) + \text{Risk}(\bar{\Theta})| \geq T_n\right).$$

We then use the empirical process to uniformly bound the stochastic residual

$$E_n(\mathcal{Z}) := L(\mathcal{Z}) - \text{Risk}(\mathcal{Z}) - L(\bar{\Theta}) + \text{Risk}(\bar{\Theta}).$$

To show this, we notice that the stochastic residual is a sum of i.i.d. r.v.'s

$$E_n(\mathcal{Z}) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{[\ell_{\mathcal{Z}}(y_i) - \ell_{\bar{\Theta}}(y_i) + \mathbb{E}\ell_{\bar{\Theta}}(y_i) - \mathbb{E}\ell_{\mathcal{Z}}(y_i)]}_{\text{mean-zero, } i.i.d. \text{ r.v.'s}}$$

$$= \frac{1}{n}\sum_{i=1}^{n} e_i + \mathrm{Risk}(\bar{\Theta}) - \mathrm{Risk}(\mathcal{Z})$$

where

$$\mathrm{Var}[e_i] \le \mathbb{E}[e_i]^{\alpha} + \frac{1}{\rho}\mathbb{E}[e_i].$$

With high probability

$$\max_{\mathcal{Z}} \frac{1}{n}\sum_{i=1}^{n} e_i \le T_n + \mathrm{Risk}(\mathcal{Z}) - \mathrm{Risk}(\bar{\Theta})$$

$$\sup_{\mathcal{Z}} E_n(\mathcal{Z}) \le T_n$$

Lemma 9. Let $\mathcal{F}$ be a class of functions, and $(y_i)_{i\in[n]}$ be an i.i.d. sample from random variable y. Suppose the variance-to-mean relationship holds uniformly over $\mathcal{F}$,

$$\mathrm{Var} f(y) = [\mathbb{E}f(y)]^{\beta} + \frac{1}{\rho}\mathbb{E}f(y), \quad \text{for all } f \in \mathcal{F},$$

where $\beta \in [0,1]$ is a constant. Then

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(y_i) \ge \mathbb{E}f(y) + T_n\right) \le \exp(-T_n).$$

To bound the right-hand side, we partition $\{\mathcal{Z} \in \mathcal{F} \colon \mathrm{Risk}(\mathcal{Z}) - \mathrm{Risk}(\bar{\Theta}) \ge L_n\}$ into a union of $A_s = \{\mathcal{Z} \in \mathcal{F} \colon 2^{s-1}L_n \le \mathrm{Risk}(\mathcal{Z}) - \mathrm{Risk}(\bar{\Theta}) < 2^s L_n\}$ for $s = 1, 2, \ldots$. Then it suffices to bounding the corresponding probability over each A_s. Towards this end, we need to bound the first and second moment of $\Delta_n(\mathcal{Z}, \mathcal{Y}_\Omega)$.

For the first moment we have

$$\inf_{\mathcal{Z}\in A_s} \mathbb{E}\Delta_n(\mathcal{Z}, \mathcal{Y}) = \inf_{\mathcal{Z}\in A_s} \mathbb{E}[L(\mathcal{Z}, \mathcal{Y}_\Omega) - L(\bar{\Theta}, \mathcal{Y}_\Omega)] \ge \inf_{\mathcal{Z}\in A_s} [\mathrm{Risk}(\mathcal{Z}) - \mathrm{Risk}(\bar{\Theta})] \ge \underbrace{2^{s-1}L_n}_{=:M(s)}$$

for any $s = 1, 2, \ldots$ For the second moment, it follows from Lemma **??** that

$$\sup_{\mathcal{Z}\in A_s} \mathrm{Var}\Delta_n(\mathcal{Z}, \mathcal{Y}) = \sup_{\mathcal{Z}\in A_s} \mathrm{Var}[L(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega) - L(\bar{\Theta}, \bar{\mathcal{Y}}_\Omega)]$$

$$\le \underbrace{M^{\frac{\alpha}{1+\alpha}}(s) + \frac{1}{\rho}M(s)}_{=:V(s)}$$

We now apply Shen & Wong for each of the set $\{\mathcal{Z} \in A_s\}$

$$\begin{aligned}
\mathbb{P}\left(\text{Risk}(\hat{\mathcal{Z}}) - \text{Risk}(\bar{\Theta}) \geq L_n\right) \leq \sum_{s=1}^{\infty} \mathbb{P}\left(\sup_{A_s}(\mathbb{E}\Delta_n - \Delta_n) \geq M_n\right) &\leq \sum_{s=1}^{\infty} \exp\left[-\frac{ncM^2(s)}{V(s) + TM(s)}\right] \qquad (6)\\
&\leq \sum_{s=1}^{\infty} \exp\left[-\frac{nM^2(s)}{M^{\frac{\alpha}{1+\alpha}}(s) + \frac{1}{\rho}M(s)}\right]\\
&\lesssim \exp(-d^{\frac{\alpha}{\alpha+1}} n^{\frac{1}{\alpha+1}})\\
&\lesssim \exp(-d), \text{ provided } n \geq d.
\end{aligned}$$

where the convergence rate $L_n > 0$ is determined by the solution to the following inequality,

$$\frac{1}{L_n} \int_{L_n}^{\sqrt{L_n^{\alpha/(\alpha+1)} + \frac{L_n}{\rho}}} \sqrt{\mathcal{H}(\varepsilon, \mathcal{F}, \|\cdot\|_2)} d\varepsilon \leq \sqrt{n}. \qquad (7)$$

In particular, the smallest L_n satisfying (7) yields the best upper bound of the error rate. Here $\mathcal{H}(\varepsilon, \mathcal{F}, \|\cdot\|_2)$ denotes the L_2-metric, ε-bracketing number (c.f. Definition 2) of family $\mathcal{F}$.

It remains to solve for the smallest possible L_n in (7). Based on Lemma 7, the inequality (7) is satisfied with

$$L_n \asymp t_n^{(\alpha+1)/(\alpha+2)} + \frac{1}{\rho} t_n, \quad \text{where } t_n = \frac{Kd_{\max}r}{n}.$$

Combining (**??**) and (6) gives

$$\text{Risk}(\hat{\mathcal{Z}}) - \text{Risk}(\bar{\Theta}) \geq \left(\frac{Kd_{\max}r}{n}\right)^{(\alpha+1)/(\alpha+2)} + \frac{1}{\rho}\left(\frac{Kd_{\max}r}{n}\right).$$

with probability at least $1 - \exp(-\sqrt{Knd_{\max}r})$..