Empirical Process: Peeling Technique

Yiping Lu
ICME, Stanford University
yplu@stanford.edu

Abstract

In this lecture note, we aim to review the peeling technique and corresponding learning rate we can achieve.

1 Problem Setting

To have an oracle error of learning algorithms in terms of a data-dependent notion of complexity, in this paper we review two paper using the localization technique

We aim to learn a function $f : \mathscr{X} \to \mathbb{R}$ from finite samples $(X_1, Y_1), \cdots, (X_n, Y_n)$ which are ndependent random variables distributed according distribution P. We define

$$P_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i), Pf = \mathbb{E}f(X), R_n f = \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(X_i)$$

Let $\sigma_1, \cdots, \sigma_n$ to be n to be n independent Rademacher random variables, that is, independent random variables for which $P(\sigma_i = 1) = P(_i= 0) = 0.5$ and the Rademencher complexity for a function class $\mathscr{F}$ is defined as $R_n\mathscr{F} = \sup_{f\in\mathscr{F}} R_n f$. The empirical (or conditional) Rademacher averages of of $\mathscr{F}$ is defined as $\mathbb{E}_\sigma R_n\mathscr{F} = \frac{1}{n}\mathbb{E}(\sup_{f\in\mathscr{F}} \sum_{i=1}^{n} \sigma_i f(X_i)|X_1, \cdots, X_n)$. In this paper, using localization technique, we can achieve a generalization bound depend on r, the fixed point of the following equation

$$r = 20\mathbb{E}_\sigma R_n\left(\{f \in \text{star}(l_f, 0) : P_n f^2 \le 2r\}\right) + \frac{13x}{n}$$

2 Main Theorems

Theorem 1 *Let $\mathscr{F}$ be a function class maps from $\mathscr{X}$ to $[a, b]$ and assume that there are some functional $T : \mathscr{F} \to \mathbb{R}$ and some constant B such that for every $f \in \mathscr{F}$, $Varf \le T(f) \le BPf$. Let ψ be a sub-root function and*

$$\psi(r) \ge \mathbb{E}R_n\{f \in \mathscr{F}, T(f) \le r\}$$

Let r^* to be the fix point of function ψ, then for every $K > 1$ with probability at least$1 - e^{-x}$ we have

$$Pf \le \frac{K}{K-1}P_n f + \frac{C_1 K}{B}r^* + \frac{x(11(b-a) + C_2 BK)}{n}$$

2.1 Technique Overview

The two main technique used in the paper is the peeling lemma and the Talagrand Concentration Inequality. In this section, we have a slightly simpler version of the proof instead of the original one in the paper.

Preprint. Under review.

Lemma 1 (Peeling Technique) *If there is a function $\phi : [0, \infty) \to [0, \infty)$ and $r^* > 0$ s.t. $\forall r > \hat{r}^*$, we have*

- $\phi(4r) \leq 2\phi(r)$
- $R_n(G_r)) \leq \phi(r)$

Then we have for all $r > \hat{r}^*$ we have

$$\mathbb{E}_{\sigma_i, z_i}\left[\frac{\frac{1}{n}\sum_{i=1}^n \sigma_i g(z_i)}{\mathbb{P}g + r}\right] \leq \frac{4\phi(r)}{r}$$

Proof: Denote $G(r)$ to be the localized set with radius r. Then we have

$$\begin{aligned}
\mathbb{E}_{\sigma_i, z_i}\left[\frac{\frac{1}{n}\sum_{i=1}^n \sigma_i g(z_i)}{\mathbb{P}g + r}\right] &\leq \sup_{g \in \mathscr{G}()} \frac{\frac{1}{n}\sum_{i=1}^n \sigma_i g(z_i)}{r} \\
&\quad + \sum_{j=0}^{\infty} \sup_{g \in \mathscr{G}(r4^{j+1}) \setminus \mathscr{G}(r4^j)} \frac{\frac{1}{n}\sum_{i=1}^n \sigma_i g(z_i)}{r4^j + r} \\
&\leq \frac{R_n(G_r)}{r} + \sum_{j=0}^{\infty} \frac{R_n(G_{r4^{j+1}+r})}{r4^j + r} \leq \frac{\phi(r)}{r} + \sum_{j=0}^{\infty} \frac{\phi(r4^{j+1} + r)}{r4^j + r} \\
&\leq \frac{\phi(r)}{r} + \sum_{j=0}^{\infty} \frac{2^{j+1}\phi(r)}{r4^j + r} \leq \frac{4\phi(r)}{r}
\end{aligned}$$

□

Theorem 2 (Talagrand Concentration Inequality) *$\mathscr{G}$ is a set of measurable functions on probability space $(Z, \mathscr{A}, P)$ and for every function in $\mathscr{G}$are bounded, mean zero and with bounded variance:*

- $\mathbb{E}[g] = 0, \mathbb{E}[g^2] \leq v, \|g\|_\infty \leq B, \forall g \in \mathscr{G}$

Then for $\forall t > 0$ we have

$$\mathbb{P}_z\left[\sup_{g \in \mathscr{T}} \frac{1}{n}\sum_{i=1}^n g(z_i) \geq 2\mathbb{E}_{z'}[\sup_{g \in \mathscr{T}} \frac{1}{n}\sum_{i=1}^n g(z_i)] + \sqrt{\frac{2tv}{n}} + \frac{2tB}{n}\right] \leq e^{-t}$$

Using Peeling technique to bound the $\mathbb{E}_{z'}[\sup_{g \in \mathscr{T}} \frac{1}{n}\sum_{i=1}^n g(z_i)]$ term in Talagrand concentration inequality, then we can get the bound we have here. The key idea behind the prove is using the ratio type concentration inequation $\frac{\frac{1}{n}\sum_{i=1}^n \sigma_i g(z_i)}{\mathbb{P}g+r}$ so that onece we do a peeling, we get a decaying exponential decay probability in each local sets and then we can apply a union bound.

2.2 Application To Empirical Risk Minimization

The main theorem can be used for bounding the generalization error for empricial risk minimization via considering the following straightforward decomposition of the generalization error

$$\mathscr{L}(f) = (P - P_n)\ell(f(x); y) + \big(P_n\ell(f(x); y) - P_n\ell(f^*(x); y)\big) + (P_n - P)\ell(f^*(x); y).$$

Strong Convexity Implies Fast Rate The condition $\text{Var} f \leq T(f) \leq BPf$ always means strongly convex loss function. Then we apply the techinque to $\underline{f - f^*}$.

Theorem 3 *For a strongly convex loss function $l : \mathscr{Y} \times \mathscr{Y} \to [0, 1]$. Define*

$$\phi_n(r) = 20\mathbb{E}_\sigma R_n\left(\{f : P_n(f - f^*)^2 \leq 2r\}\right) + \frac{13x}{n}$$

and $\hat{r}^*$ be the fixed point of function ϕ_n. Then for the minimizer $\hat{f} = \arg\min_{f\in\mathscr{F}} P_n l_f$ with probablity at least $1 - e^{-x}$ we have

$$Pl_{\hat{f}} < L^* + c(\hat{r}^* + \frac{x}{n})$$

Here $L^* = \inf_{f\in\mathscr{F}} Pl_f$

Slow rate without strong convexity When the loss function is not strongly convex, we can still meet Var$f \le T(f) \le BPf$ via using the boundness of the the loss function. In this case, the rate we get is

Theorem 4 *For a loss function $l : \mathscr{Y} \times \mathscr{Y} \to [0, 1]$. Define*

$$\phi_n(r) = 20\mathbb{E}_\sigma R_n\left(\{f \in \textit{star}(l_f, 0) : P_n f^2 \le 2r\}\right) + \frac{13x}{n}$$

and $\hat{r}^*$ be the fixed point of function ϕ_n. Then for the minimizer $\hat{f} = \arg\min_{f\in\mathscr{F}} P_n l_f$ we have

$$Pl_{\hat{f}} < L^* + c(\sqrt{L^* r^*} + r^*)$$

Here $L^* = \inf_{f\in\mathscr{F}} Pl_f$

3 Examples

3.1 VC Classes

Our next example considers VC-type classes. Although this classical example has been extensively studied in learning theory, our results provide strict improvements over antecedents.

One general definition of VC-type classes (which is not necessarily binary) uses the metric entropy condition. Consider a loss class $l \circ \mathscr{H}$ that satisfies

$$\log \mathscr{N}(\varepsilon, l \circ \mathscr{H}, L_2(P_n)) \le O\left(d \log \frac{1}{\epsilon}\right),$$

where d is th so-called the Vapnik–Chervonenkis (VC) dimension. Using Dudley's integral bound to find the surrogate ψ and solving $r \le O(\psi(r; \delta))$, it can be proven that

$$r^* \le O\left(\frac{d \log n}{n}\right).$$

Note: we get $\log(n)/n$ rate which **is faster than the $1/\sqrt{n}$ rate we get in the class.**

Application We can have $\frac{d\log(n)}{n}$ rate for learning linear regression on bounded features.

3.2 Complexity Assumptions

Next we consider the function space with complexity assumption

$$\log \mathscr{N}(\varepsilon, l \circ \mathscr{H}, L_2(P_n)) \le O(\epsilon^{-2\rho})$$

Then using duely integral we know $R_n(G(r)) \le \frac{r^{\frac{1-\rho}{2}}}{\sqrt{n}}$, then the solution to the fixed point equation is $\hat{r} = n^{-\frac{1}{1+\rho}}$.

Comparison with $n\delta_n^2 = \log \mathscr{N}(\delta_n, \mathscr{F}, \|\cdot\|_2)$. Notice that the radius we select in local radamecher complexity is $r^* = \delta_n^2$. Thus we exactly get the same rate as the traitional nonparametric statistics get using $n\delta_n^2 = \log \mathscr{N}(\delta_n, \mathscr{F}, \|\cdot\|_2)$ and get δ_n^2 rate when the loss is strongly convex.

3.3 Kernel Classifiers

Denote $(\hat{\lambda}_i)_{i=1}^n$ as the eigenvalue of the normalized Gram matrix $\hat{T} = \frac{1}{n}(k(X_i, X_j))_{i,j=1,\cdots,n}$, then the local radamecher complexity is $\mathbb{E}R_n(G(r)) = \left(\frac{2}{n}\sum_{i=1}^{\infty}\min\{r, \lambda_i\}\right)^{1/2}$. If we have a assumption on the polynoimal decay on $(\hat{\lambda}_i)_{i=1}^n$, then with optimal selection of r we can get min-max optimal rates for kernel classifiers.

3.4 Neural Networks

Let's consider the function space of Relu neural network.

Lemma 2 *For the function space*

$$\Phi(L, W, S, B) := \{(W^{(L)}\eta(\cdot) + b^{(L)}) \circ \cdots \circ (W^{(1)}\eta(\cdot) + b^{(1)})$$
$$|W^{(l)} \in \mathbb{R}^{W\times W}, b^{(l)} \in \mathbb{R}^W, \sum_{l=1}^{L}(\|W^l\|_0 + \|b^l\|_0) \leq S, \max_l \|W^l\|_\infty + \|b^(l)\|_\infty = B\}$$

We have the covering number bound

$$\log N(\epsilon, \mathscr{F}_2, \|\cdot\|_\infty) \leq O(SL\log(\frac{LB(W+1)}{\epsilon}))$$

To estimate the local rademacher complexity of a deep neural network, we always use the duley integral

$$\begin{aligned}\mathbb{E}[\hat{R}_n(G_r)] &\leq \mathbb{E}\left[\inf_{\alpha>0}\{4\alpha + \frac{\sqrt{n}}{12}\int_\alpha^{\sqrt{2r/\alpha}}\sqrt{\log 2\mathscr{N}(\epsilon, \Phi(L, W, S, B), \|\cdot\|_{n,2})}d\epsilon\}\right] \\ &\leq \frac{1}{n} + \frac{1}{\sqrt{n}}\int_{1/n}^{\sqrt{2r/\alpha}}\sqrt{CSSL\log(LB(W+1)\epsilon^{-1})}d\epsilon \\ &\leq \sqrt{\frac{SLr}{n}\log(L(B)(W+1)n)}\end{aligned}$$

Applications Using the local rademencher complexity with approximation power of the neural network, we can obtain oracle rate for learning Neural network in many function spaces:

- Nonparametric learning deep Relu NN in Holder space[3]: $N^{-p/d} + \frac{N(\log(N)+\log(n))}{n} + \frac{1}{n}$
- Nonparametric learning deep Relu NN in Besov space [4]:$N^{-2s/d} + \frac{N(\log(N)+\log(n))}{n} + \frac{1}{n}$

Here N is the size of the Neural Network. With the optimal selection of N, we can achieve the min-max rate of using Neural network for non-parametric learning using ERM.

4 Limitations and future work

Local radamecher complexity can't deal with r regression with square loss and general classes of functions without the boundedness assumptions. To achieve the fast rate, the strong convexity assumption and the lipschitz condition to have contraction inequatily make the loss function to be bounded. This makes localization techinques hard to be used in heavy-tail cases and robustness setting.

It's also interesting to consider how to use the localization technique for the interpolation estimators with implicit regularisation.

5 Setting for Optimal Problem Dependent Generalization Error Bounds

This paper aims to provide generalization errors that scale near-optimally with the variance, the effective loss, or the gradient norms evaluated at the "best hypothesis.", *i.e.* we get learning rate depend on

$$\mathscr{V}^* := \text{Var}[\ell(h^*; z)], \quad \mathscr{L}^* := P[\ell(h^*; z) - \inf_H \ell(h; z)].$$

In the previous paper review, we already get

Statement 1 (current blueprint) *Assume that ψ is a sub-root function, i.e., $\psi(r;\delta)/\sqrt{r}$ is non-increasing with respect to $r \in \mathbb{R}_+$. Assume the Bernstein condition $T(f) \leq B_e Pf$, $B_e > 0$, $\forall f \in F$. Then with probability at least $1-\delta$, for all $f \in F$ and $K > 1$,*

$$(P - P_n)f \leq \frac{1}{K}Pf + \frac{C(K-1)r^*}{B_e},$$

where r^* is the "fixed point" solution of the equation $r = B_e\psi(r;\delta)$.

Statement 1 has become a standard tool in learning theory. However, it requires a rather technical proof, and it appears to be loose when compared with the original assumption

$$\sup_{f \in F: T(f) \leq r} (P - P_n)f \leq \psi(r;\delta). \tag{1}$$

In this section, we would like to directly extend (1) to hold uniformly without sacrificing any accuracy.

6 Main Theorem

6.1 Technique Overview:uniform localized convergence

The key intuition behind this paper is that the uniform restatement of the "localized" argument (1) is nearly cost-free, because the deviations $(P - P_n)g_f$ can be controlled solely by the real valued functional $T(f)$. The intuition can be shown in the following lemma

Lemma 3 (the "uniform localized convergence" argument) *For a function class $G = \{g_f : f \in F\}$ and functional $T : F \to [0, R]$, assume there is a function $\psi(r;\delta)$, which is non-decreasing with respect to r and satisfies that $\forall \delta \in (0,1)$, $\forall r \in [0, R]$, with probability at least $1-\delta$,*

$$\sup_{f \in F: T(f) \leq r} (P - P_n)g_f \leq \psi(r;\delta). \tag{2}$$

Then, given any $\delta \in (0,1)$ and $r_0 \in (0, R]$, with probability at least $1-\delta$, for all $f \in F$, either $T(f) \leq r_0$ or

$$(P - P_n)g_f \leq \psi\left(2T(f); \delta\left(\log_2 \frac{2R}{r_0}\right)^{-1}\right). \tag{3}$$

Proof: To formalize the idea that the deviations $(P - P_n)g_f$ can be controlled solely by the real valued functional $T(f)$. The author apply a "peeling" technique: we take $r_k = 2^k r_0$, where $k = 1, 2, \ldots, \lceil \log_2 \frac{R}{r_0} \rceil$ and then apply a union bound to extend (1) to hold for all r_k.

For any $f \in F$ such that $T(f) > r_0$ is true, there exists a non-negative integer k such that $2^k r_0 < T(f) \leq 2^{k+1} r_0$. Then we have

$$(P - P_n)g_f \leq \psi\left(r_{k+1}; \delta\left(\log_2 \frac{2R}{r_0}\right)^{-1}\right) \leq \psi\left(2T(f); \delta\left(\log_2 \frac{2R}{r_0}\right)^{-1}\right),$$

□

Slow rate Regime Taking optimal choice of K in Statement 1, we can re-write the conclusion as

$$(P - P_n)f \leq 20\sqrt{\frac{r^* Pf}{B_e}} - \frac{r^*}{B_e}.$$

where the right hand side is of order $\sqrt{r^* Pf / B_e}$ when $Pf < r^*/B_e$, and order r^*/B_e when $Pf \leq r^*/B_e$.

At the same time,

$$\psi(2T(f);\delta) \quad \leq \quad \psi(2B_e Pf;\delta) \leq \frac{\sqrt{2B_e Pf}}{\sqrt{r^*}}\psi(r^*;\delta) \leq \sqrt{\frac{2r^* Pf}{B_e}}. \tag{4}$$

This means that the "uniform localized convergence" argument (3) strictly improves over Statement 1 (ignoring negligible $O(\log\log n)$ factors).

Fast rate Regime In the fast rate regime, the removal of the "sub-root" requirement on ψ allows one to achieve parameter localization which added flexibility in the choice of one-sided uniform inequalities and uniform convergence of gradient vectors. This bring us several benefits

- In traditional analysis, a loose "sub-root" surrogate function is often obtained via two-sided concentration and Lipchitz contraction, making global Lipchitz constants unavoidable.
- The removal of the "sub-root" restriction is crucial because under curvature and smoothness conditions, the uniform error of excess loss typically grows "faster" than the square root function.
- Simple "truncated" functions can be used to established one-sided uniform inequalities that are sharper than two-sided ones, which enable recovery of results in unbounded and heavy-tailed regression problems.

Principle of uniform localized convergence. First, determine the concentrated functions, the measurement functional and the surrogate ψ, and obtain a sharp "uniform localized convergence" argument. Then, perform localization analysis that is customized to the problem setting and the learning algorithm. Distinct from the blueprint, the right hand side of "uniform localized convergence" argument (3) contains a "free" variable $T(f)$ rather than a fixed value r^*.

6.2 Loss-dependent rates via empirical risk minimization

Theorem 5 (loss-dependent rate of ERM) *For the excess loss class F in (**??**), assume there is a meaningful surrogate function $\psi(r;\delta)$ that satisfies $\forall\delta \in (0,1)$ and $\forall r > 0$, with probability at least $1-\delta$,*

$$\sup_{f\in F: P[f^2]\leq r} (P - P_n)f \leq \psi(r;\delta).$$

Then the empirical risk minimizer $\hat{h}_{\text{ERM}} \in \arg\min_H\{P_n\ell(h;z)\}$ satisfies for any fixed $\delta \in (0,1)$ and $r_0 \in (0, 4B^2)$, with probability at least $1-\delta$,

$$\mathscr{L}(\hat{h}_{\text{ERM}}) \leq \psi\left(24B\mathscr{L}^*; \frac{\delta}{C_{r_0}}\right) \vee \frac{r^*}{6B} \vee \frac{r_0}{48B},$$

where $C_{r_0} = 2\log_2 \frac{8B^2}{r_0}$, and r^* is the fixed point of $6B\psi\left(8r; \frac{\delta}{C_{r_0}}\right)$.

6.3 Variance-dependent rates via moment penalization

In this section, this paper consider the following algorithm:

- At the first-stage, we derive a preliminary estimate of $L_0^* := P\ell(h^*;z)$ via the "auxiliary" data set S', which we refer to as $\widehat{L_0^*}$. Then, at the second stage, we perform regularized empirical risk minimization on the "primal" data set S, which penalizes the centered second moment $P_n[(\ell(h;z) - \widehat{L_0^*})^2]$.

- Let $\psi(r;\delta)$ be a meaningful surrogate function that satisfies $\forall\delta \in (0,1)$, $\forall r > 0$, with probability at least $1-\delta$,

$$4\mathfrak{R}_n\{f \in F : P_n[f^2] \le 2r\} + \sqrt{\frac{2r\log\frac{8}{\delta}}{n}} + \frac{9B\log\frac{8}{\delta}}{n} \le \psi(r;\delta).$$

Denote $C_n = 2\log_2 n + 5$. Given a fixed $\delta \in (0,1)$, let the estimator $\hat{h}_{\text{MP}}$ be

$$\hat{h}_{\text{MP}} \in \arg\min_H \left\{ P_n\ell(h;z) + \psi\left(16P_n[(\ell(h;z) - \widehat{L_0^*})^2]; \frac{\delta}{C_n}\right)\right\}.$$

Then we have

Theorem 6 (variance-dependent rate) *Let $\widehat{L_0^*} = \inf_H \mathbb{P}_{S'}\ell(h;z)$ be attained via empirical risk minimization on the auxiliary data set S'. Assume that the meaningful surrogate function $\psi(r;\delta)$ is "sub-root," i.e. $\frac{\psi(r;\delta)}{\sqrt{r}}$ is non-increasing over $r \in [0, 4B^2]$ for all fixed δ. Then for any $\delta \in (0, \frac{1}{2})$, by performing the moment-penalized estimator, with probability at least $1-2\delta$,*

$$\mathscr{E}(\hat{h}_{\text{MP}}) \le 2\psi\left(c_1V^*; \frac{\delta}{C_n}\right) \vee \frac{c_1r^*}{8B},$$

where r^* is the fixed point of $B\psi(r; \frac{\delta}{C_n})$ and c_1 is an absolute constant.

Comparison with existing results. The best variance-dependent rate attained by existing estimators is of the order

$$\sqrt{\frac{V^*r^*}{B^2}} \vee \frac{r^*}{B},$$

which is strictly worse than the rate proved in Theorem 6. The reasoning is similar to what we shown before: the bound can perform much better when $V^* \ge \Omega(r^*)$ for

$$\psi(V^*;\delta) \overset{\text{sub-root}}{\le} \sqrt{\frac{V^*}{r^*}}\psi(r^*;\delta) \overset{\text{fixed point}}{=} O\left(\sqrt{\frac{V^*r^*}{B^2}}\right).$$

7 Examples

VC Classes For the VC classes satisfies

$$\log \mathscr{N}(\varepsilon, l\circ\mathscr{H}, L_2(P_n)) \le O\left(d\log\frac{1}{\epsilon}\right),$$

where d is th so-called the Vapnik–Chervonenkis (VC) dimension. Using Dudley's integral bound to find the surrogate ψ and solving $r \le O(\psi(r;\delta))$, it can be proven that

$$\psi(r;\delta) \le O\left(\sqrt{\frac{dr}{n}\log\frac{8B^2}{r}} \vee \frac{Bd}{n}\log\frac{8B^2}{r}\right), \quad r^* \le O\left(\frac{B^2d\log n}{n}\right).$$

Thus we can get the result

$$\mathscr{E}(\hat{h}_{\text{MP}}) \le O\left(\sqrt{\frac{dV^*\log\frac{8B^2}{V^*}}{n}} \vee \frac{Bd\log n}{n}\right). \tag{5}$$

The result matches the $\Omega(\sqrt{\frac{dV^*}{n}})$ lower bound and closes the $O(\log n)$ gap in the regime $V^* \ge \Omega(\frac{B^2}{(\log n)^\alpha})$ in the previous results.

Non-parametric classes of polynomial growth Consider the metric entropy condition

$$\log \mathscr{N}(\varepsilon, \ell \circ H, \|\cdot\|_{n,2}) \leq O\left(\varepsilon^{-2\rho}\right), \rho \in (0,1) \tag{6}$$

Using Dudley's integral we can verify that

$$\psi(r;\delta) \leq O\left(\sqrt{\frac{r^{1-\rho}}{n}}\right), \quad r^* \leq O\left(\frac{B^{\frac{2}{1+\rho}}}{n^{\frac{1}{1+\rho}}}\right).$$

As a result, the bound using the technique here can get the order

$$\mathscr{E}(\hat{h}_{\text{MP}}) \leq O\left(V^{*\frac{1-\rho}{2}} n^{-\frac{1}{2}} \vee \frac{r^*}{B}\right), \tag{7}$$

which is $O\left(V^{*\frac{1-\rho}{2}} n^{-\frac{1}{2}}\right)$ when $V^* \geq \Omega(r^*)$.

Compared with the previous results

$$\mathscr{E}(\hat{h}_{\text{previous}}) \leq O\left(\sqrt{V^*} B^{-\frac{\rho}{1+\rho}} n^{-\frac{1}{2+2\rho}} \vee \frac{r^*}{B}\right), \tag{8}$$

We consider the improvement of the algorithm in the following two regimes

The "traditional" regime. The more "traditional" regime: $B \approx 1$, $V^* \approx n^{-a}$ where $a > 0$ is a fixed constant. The improvement is $1 \vee (V^* n^{\frac{1}{1+\rho}})^{\frac{\rho}{2}}$. If $V^* \approx n^{-a}$ where $0 < a < \frac{1}{1+\rho}$, the variance-dependent rate improves by orders polynomial in n.

The "high-risk" regime. $B \approx n^b$ where $b > 0$ is a fixed constant, and $V^* \ll B^2$ (i.e., V^* is much smaller than order n^{2b}). Under the simple situation $B^{\frac{2}{1+\rho}} \leq V^* \ll 4B^2$, an improvement of order $O(n^{\frac{\rho}{2}(\frac{1}{1+\rho})})$ relative to the previous result has been achieved. By letting $\rho \to 1$ our improvement can be as large as $O(n^{\frac{1}{4}})$ and the larger ρ, the more improvement can be provided.

Remark. The "high-risk" regime captures modern contains problems such as counterfactual risk minimization, policy learning, and supervised learning with limited number of samples which are "high-risk" learning problems.

Reference

[1] Xu Y, Zeevi A. Towards Optimal Problem Dependent Generalization Error Bounds in Statistical Learning Theory. arXiv preprint arXiv:2011.06186, 2020.

[2] Foster D J, Syrgkanis V. Orthogonal statistical learning. arXiv preprint arXiv:1901.09036, 2019.

[3] J. Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. ArXiv e-prints, August 2017.

[4] Suzuki T. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality[J]. arXiv preprint arXiv:1810.08033, 2018.