Exact Gap between Generalization Error and Uniform Convergence in Random Feature Models(2021, arxiv)

SeongSik Choi

January 26, 2022

Seoul National University

Motivation

- 학습 이론에서, test err와 train err의 차이가 작을수록 generalize 성능이 좋다는 의미.
- Uniform convergence(UC) is the supremum of (test err - train err) over a certain function class.
- $\therefore$ UC is a powerful tool for understanding the generalization performance of predictors.
- Bounds on uniform convergence는 일반적으로 $\sqrt{\text{complexity}/n}$의 형태이므로 만약 tight하다면 function class가 복잡할수록 generalize를 못 해야한다. (그러나 현실은 DNN도 generalize를 잘 한다.)
- **Question :** What is the exact gap between uniform convergence and the actual generalization errors?

Model setup

- Consider a dataset $(\boldsymbol{x}_i, y_i)_{i\in[n]}$ with n samples.
- The covariates follow $\boldsymbol{x}_i \sim_{iid}$ Unif $\left(\mathbb{S}^{d-1}(\sqrt{d})\right)$.
- Responses satisfy $y_i = f_d\left(\boldsymbol{x}_i\right) + \varepsilon_i$.
- The noises satisfying $\varepsilon_i \sim_{iid} \mathcal{N}\left(0, \tau^2\right)$ indep of $(x_i)_{i\in[n]}$.
- The noisy setting $\left(\tau^2 > 0\right)$ and noiseless setting $\left(\tau^2 = 0\right)$.
- Function class : Random features(RF) function class.

Random features model

(Random features model) Let $(\boldsymbol{\theta}_j)_{j\in[N]} \sim_{iid} \text{Unif}\left(\mathbb{S}^{d-1}(\sqrt{d})\right)$ be the random feature vectors. Given an activation function $\sigma : \mathbb{R} \to \mathbb{R}$, we define the random features function class $\mathcal{F}_{\text{RF}}(\boldsymbol{\Theta})$ by

$$\mathcal{F}_{\text{RF}}(\boldsymbol{\Theta}) \equiv \left\{ f_{\boldsymbol{a}}(\boldsymbol{x}) = \sum_{j=1}^{N} a_j \sigma\left(\langle \boldsymbol{x}, \boldsymbol{\theta}_j \rangle / \sqrt{d}\right) : \boldsymbol{a} \in \mathbb{R}^N \right\}$$

$$\min_{f_{\boldsymbol{a}} \in \mathcal{F}_{\text{RF}}(\boldsymbol{\Theta})} R(f_{\boldsymbol{a}}) = \min_{\boldsymbol{a}\in\mathbb{R}^N} R(\boldsymbol{a})$$

This function class is proper because it can be interpreted as a linearized version of two layer neural networks.

Generalization error

(Generalization error) We define population risk and the empirical risk of a predictor $\boldsymbol{a} \in \mathbb{R}^N$ as

$$R(\boldsymbol{a}) = \mathbb{E}_{\boldsymbol{x},y} \left(y - f_{\boldsymbol{a}}(\boldsymbol{x})\right)^2$$

$$\widehat{R}_n(\boldsymbol{a}) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - f_{\boldsymbol{a}}(\boldsymbol{x_i})\right)^2$$

and the empirical risk minimizer with vanishing regularization as

$$\boldsymbol{a}_{\min} = \lim_{\lambda \to 0+} \arg\min_{\boldsymbol{a}} \left[\widehat{R}_n(\boldsymbol{a}) + \lambda \|\boldsymbol{a}\|_2^2\right].$$

In the overparameterized regime ($N > n$), under mild conditions, we have $\min_{\boldsymbol{a}} \widehat{R}_n(\boldsymbol{a}) = \widehat{R}_n\left(\boldsymbol{a}_{\min}\right) = 0$. In this regime, $\boldsymbol{a}_{\min}$ can be interpreted as the minimum ℓ_2 norm interpolator.

$\Rightarrow$ $R(\boldsymbol{a}_{\min})$: Generalization error

Uniform convergence bounds

(Uniform convergence bounds) We denote the uniform convergence bound over a norm ball and the uniform convergence over interpolators in the norm ball by

$$U(A, N, n, d) \equiv \sup_{(N/d)\|\boldsymbol{a}\|_2^2 \leq A} \left(R(\boldsymbol{a}) - \widehat{R}_n(\boldsymbol{a}) \right)$$

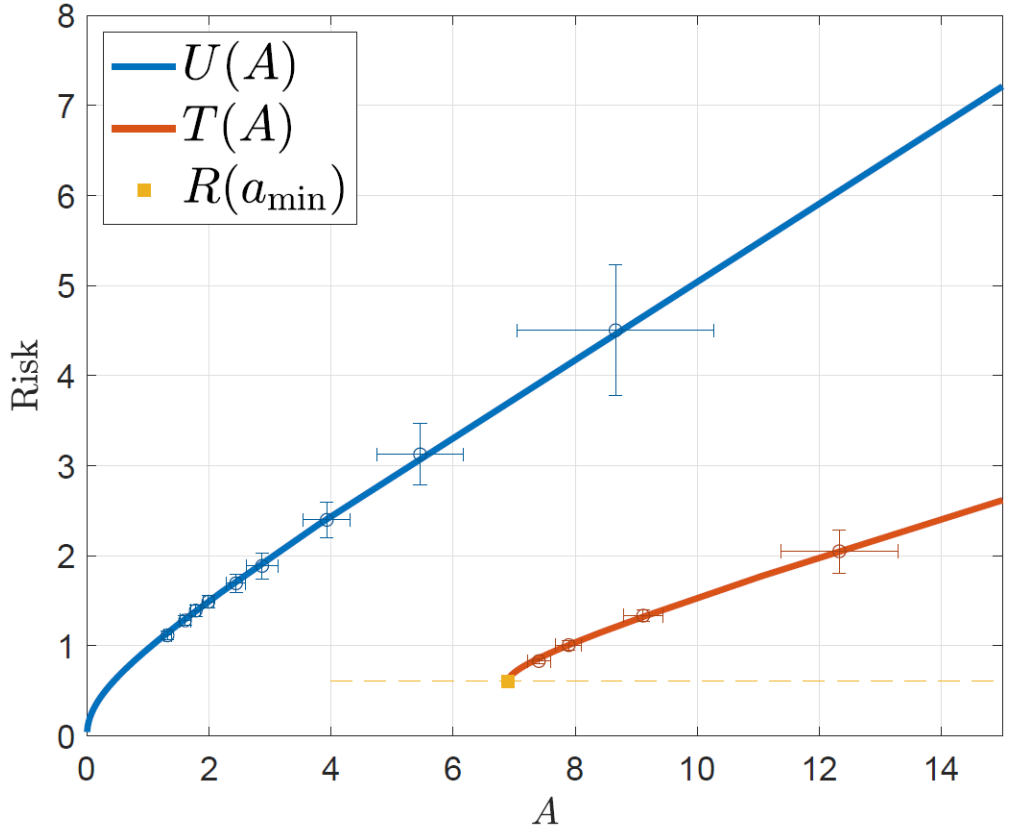
$$T(A, N, n, d) \equiv \sup_{(N/d)\|\boldsymbol{a}\|_2^2 \leq A, \widehat{R}_n(\boldsymbol{a})=0} R(\boldsymbol{a})$$

Note that we need $A \geq (N/d) \left\|\boldsymbol{a}_{\min}\right\|_2^2$ to satisfy $\{\boldsymbol{a} \in \mathbb{R}^N : (N/d)\|\boldsymbol{a}\|_2^2 \leq A, \widehat{R}_n(\boldsymbol{a}) = 0\} \neq \phi$.

We want to quantify $U(A), T(A), R(\boldsymbol{a}_{\min})$ (abuse of notation)

Uniform convergence bounds and Generalization error

We want to quantify $U(A), T(A), R(\boldsymbol{a}_{\min})$ (abuse of notation)



σ : Relu, $(N, n, d) = (500, 300, 200)$

$U(A) \geq T(A) \geq R(\boldsymbol{a}_{\min})$ for any $A \geq (N/d) \|\boldsymbol{a}_{\min}\|_2^2$

Asymptotic limit ($d \to \infty$)

We approach this problem in the limit $d \to \infty$ with $N/d \to \psi_1$(the number of parameters) and $n/d \to \psi_2$(sample size). We further assume the setting of a linear target function f_d and a nonlinear activation function σ.

In this regime, our main result Theorem 1 will show that,

$$U(A, N, n, d) \overset{d\to\infty}{\to} \mathcal{U}(A, \psi_1, \psi_2),$$
$$T(A, N, n, d) \overset{d\to\infty}{\to} \mathcal{T}(A, \psi_1, \psi_2),$$

and Theorem 1 of Mei & Montanari (2019) shows

$$(N/d)\,\|\boldsymbol{a}_{\min}\|_2^2 \overset{d\to\infty}{\to} \mathcal{A}(\psi_1, \psi_2),$$
$$R(\boldsymbol{a}_{\min}) \overset{d\to\infty}{\to} \mathcal{R}(\psi_1, \psi_2).$$

Asymptotic limit ($d \to \infty$ + $\psi_1 \to \infty$)

In the limit $\psi_1 \to \infty$ after $d \to \infty$, we denote

$$\mathcal{U}_\infty(A, \psi_2) \equiv \lim_{\psi_1 \to \infty} \mathcal{U}(A, \psi_1, \psi_2)$$

$$\mathcal{T}_\infty(A, \psi_2) \equiv \lim_{\psi_1 \to \infty} \mathcal{T}(A, \psi_1, \psi_2)$$

$$\mathcal{A}_\infty(\psi_2) \equiv \lim_{\psi_1 \to \infty} \mathcal{A}(\psi_1, \psi_2)$$

$$\mathcal{R}_\infty(\psi_2) \equiv \lim_{\psi_1 \to \infty} \mathcal{R}(\psi_1, \psi_2)$$

Asymptotic limit (substituting A)

In order for $\mathcal{U}$ and $\mathcal{T}$ to serve as proper bounds for $R(\boldsymbol{a}_{\min})$, we need to take at least $A \geq \psi_1 \|\boldsymbol{a}_{\min}\|_2^2$. Therefore, we will choose $A = \alpha\psi_1 \|\boldsymbol{a}_{\min}\|_2^2$ for some $\alpha > 1$.

Define

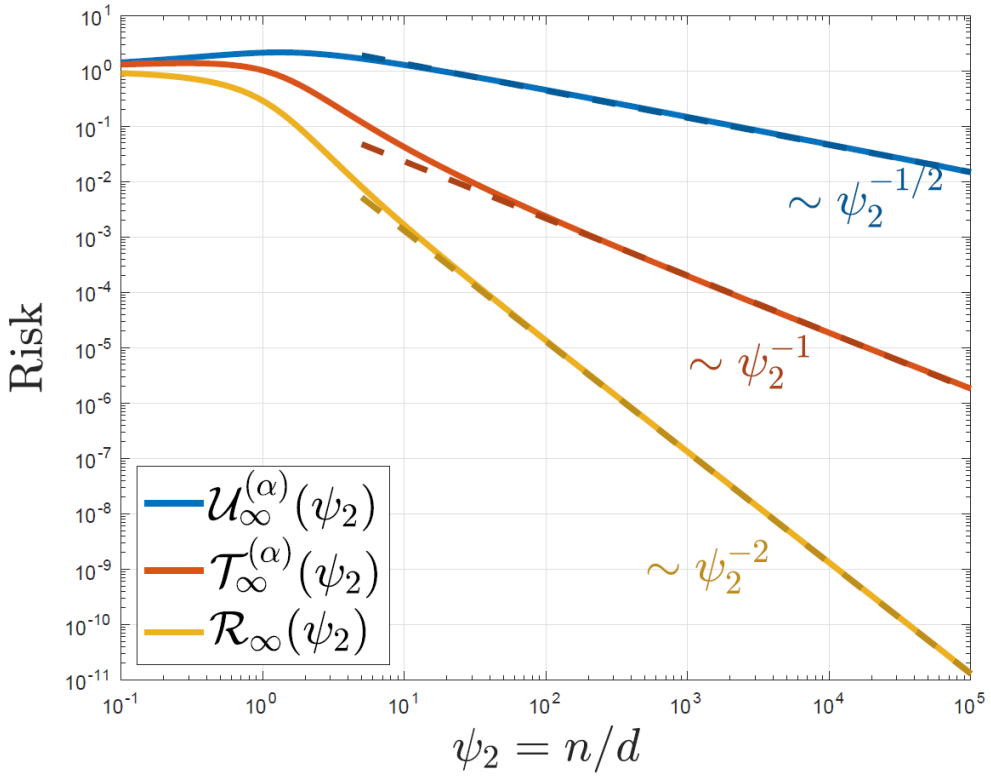
$$\mathcal{U}^{(\alpha)}(\psi_1, \psi_2) \equiv \mathcal{U}(\alpha\mathcal{A}(\psi_1, \psi_2), \psi_1, \psi_2),$$
$$\mathcal{T}^{(\alpha)}(\psi_1, \psi_2) \equiv \mathcal{T}(\alpha\mathcal{A}(\psi_1, \psi_2), \psi_1, \psi_2),$$

and

$$\mathcal{U}_\infty^{(\alpha)}(\psi_2) \equiv \lim_{\psi_1 \to \infty} \mathcal{U}^{(\alpha)}(\psi_1, \psi_2),$$
$$\mathcal{T}_\infty^{(\alpha)}(\psi_2) \equiv \lim_{\psi_1 \to \infty} \mathcal{T}^{(\alpha)}(\psi_1, \psi_2).$$

Asymptotic power laws ($d \to \infty$ + $\mathcal{A}_\infty$ + noiseless)

$$\mathcal{U}_\infty^{(\alpha)}\left(\psi_2; \tau^2 = 0\right) \sim \psi_2^{-1/2}$$
$$\mathcal{T}_\infty^{(\alpha)}\left(\psi_2; \tau^2 = 0\right) \sim \psi_2^{-1}$$
$$\mathcal{R}_\infty^{(\alpha)}\left(\psi_2; \tau^2 = 0\right) \sim \psi_2^{-2}$$



Asymptotic power laws ($d \to \infty$ + $\mathcal{A}_\infty$ + noisy)

$$\mathcal{U}_\infty^{(\alpha)}\left(\psi_2;\tau^2\right) - \tau^2 \sim \psi_2^{1/2}$$
$$\mathcal{T}_\infty^{(\alpha)}\left(\psi_2;\tau^2\right) - \tau^2 \sim 1$$
$$\mathcal{R}_\infty\left(\psi_2;\tau^2\right) - \tau^2 \sim \psi_2^{-1}$$

In the presence of label noise, the excess risk $\mathcal{R}_\infty - \tau^2$ vanishes in the large sample limit. In contrast, the classical uniform convergence $\mathcal{U}_\infty$ becomes vacuous, whereas the uniform convergence over interpolators $\mathcal{T}_\infty$ converges to a constant, which gives a non-vacuous bound of $\mathcal{R}_\infty$.

Asymptotic power laws ($d \to \infty$)

The inferred asymptotic law gives (c.f. Figure 4)

$$\mathcal{U}^{(\alpha)}(\psi_1, \psi_2) - \mathcal{U}_{\infty}^{(\alpha)}(\psi_2) \sim \psi_1^{-1}$$
$$\mathcal{T}^{(\alpha)}(\psi_1, \psi_2) - \mathcal{T}_{\infty}^{(\alpha)}(\psi_2) \sim \psi_1^{-1}$$
$$\mathcal{R}(\psi_1, \psi_2) - \mathcal{R}_{\infty}(\psi_2) \sim \psi_1^{-1}$$

Note that large ψ_1 should be interpreted as the model being heavily overparametrized (a large width network). This asymptotic power law implies that, both uniform convergence bounds correctly predict the decay of the test error with the increase of the number of features.

Conclusion

- They focus on investigating uniform convergence bound(U), uniform convergence bound over interpolators(T) and generalization error(R).
- The contribution of this paper is that they provide that the asymptotic limit of U, T that can be calculated through proof.
- They conjecture asymptotic power laws using simulations(figure).
- Their results exhibit a setting in which standard uniform convergence bound is vacuous while uniform convergence over interpolators gives a non-trivial bound of the actual generalization error.

Reference

- Yang, Z., Bai, Y., Mei, S. (2021). Exact gap between generalization error and uniform convergence in random feature models. arXiv preprint arXiv:2103.04554.

- Mei, S., Montanari, A. (2019). The Generalization Error of Random Features Regression: Precise Asymptotics and the Double Descent Curve. Communications on Pure and Applied Mathematics.