Exact Gap between Generalization Error and Uniform Convergence*

Speaker: Zitong Yang

Yu Bai

Song Mei

ICML 2021

* https://arxiv.org/pdf/2103.04554.pdf

Uniform Convergence Puzzle

- Given a training set of size n

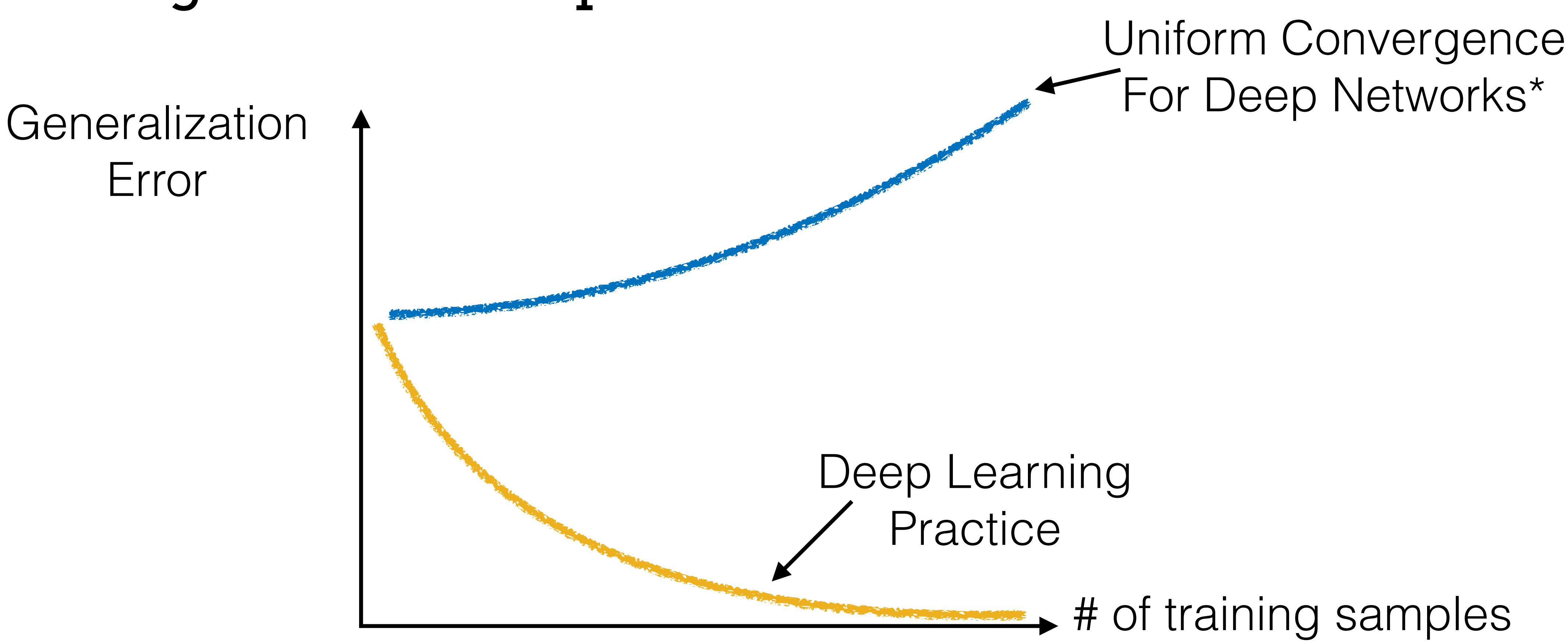
$$\underbrace{R(f) - \hat{R}_n(f)}_{\text{Generalization Error}} \leq \overbrace{\sup_{f \in \mathcal{F}} \left\{ R(f) - \hat{R}_n(f) \right\}}^{\text{Uniform Convergence}} = O\left(\sqrt{\frac{\text{Complexity of } \mathcal{F}}{n}} \right)$$

Uniform Convergence Puzzle

- **Given a training set of size n**

$$\underbrace{R(f) - \hat{R}_n(f)}_{\text{Generalization Error}} \le \overbrace{\sup_{f\in\mathscr{F}} \left\{ R(f) - \hat{R}_n(f) \right\}}^{\text{Uniform Convergence}} = O\left(\sqrt{\frac{\text{Complexity of } \mathscr{F}}{n}}\right)$$

- **The generalization puzzle**



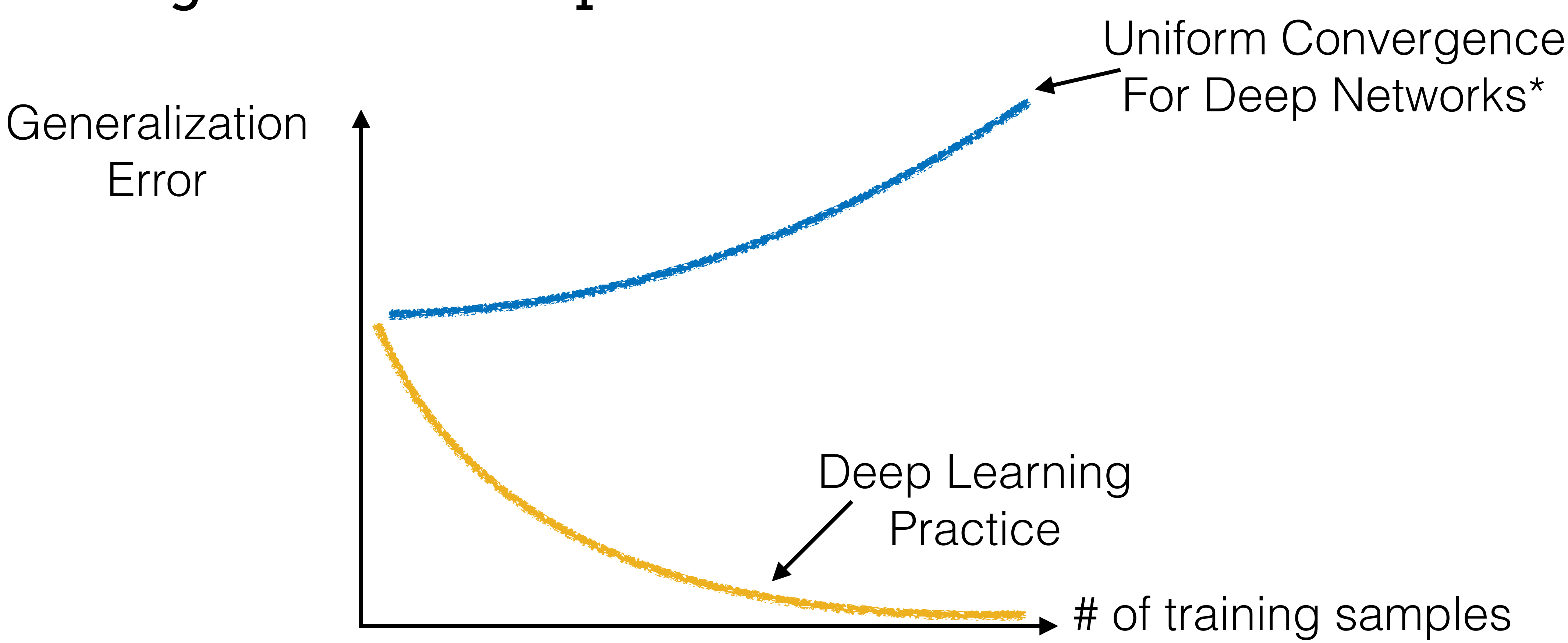
** Nagarajan&Kolter, Uniform convergence may be unable to explain generalization*

Solution: Capacity Reduced Model Class

- **Fix some $\mathscr{F}_{\text{reduced}} \subset \mathscr{F}$**

$$\underbrace{R(f) - \hat{R}_n(f)}_{\text{Generalization Error}} \leq \overbrace{\sup_{f \in \mathscr{F}_{\text{reduced}}} \left\{ R(f) - \hat{R}_n(f) \right\}}^{\text{Uniform Convergence}} = O\left(\sqrt{\frac{\text{Complexity of } \mathscr{F}_{\text{reduced}}}{n}} \right)$$

- **The generalization puzzle**



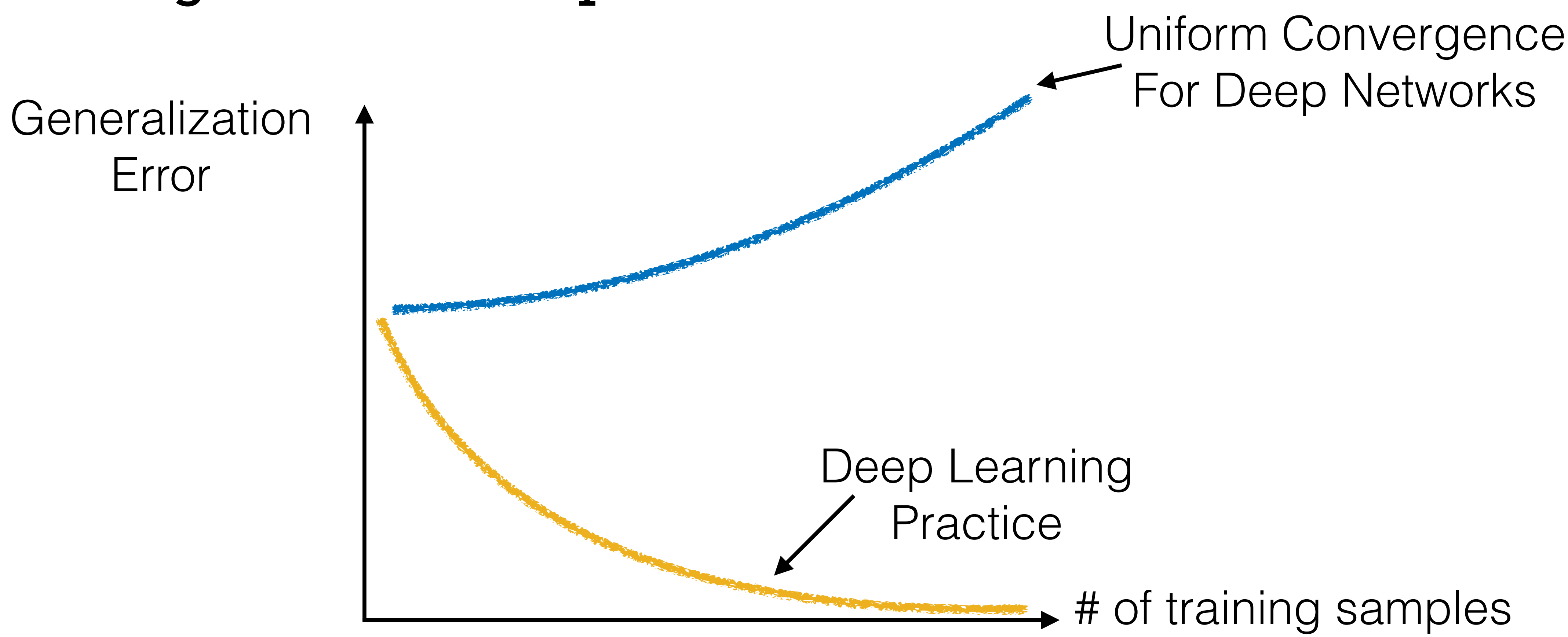
** Nagarajan&Kolter, Uniform convergence may be unable to explain generalization*

Solution: Capacity Reduced Model Class

- As deep networks interpolates, a natural class to consider is*

$$\mathscr{F}_{\text{reduced}} = \left\{ f \in \mathscr{F}, \hat{R}_n(f) = 0 \right\} \subset \mathscr{F}$$

- The generalization puzzle



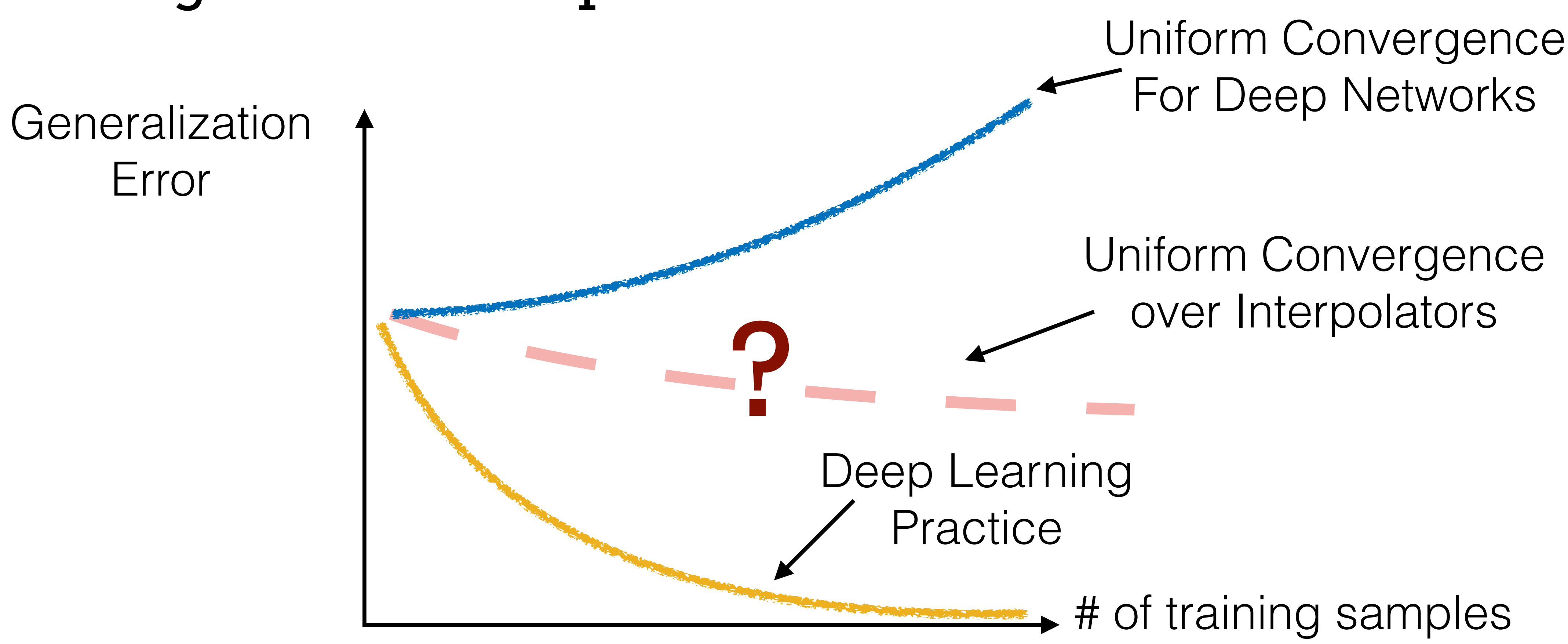
* Zhou, Sutherland, Srebro. On uniform convergence and low-norm interpolation learning.

Solution: Capacity Reduced Model Class

- As deep networks interpolates, a natural class to consider is*

$$\mathscr{F}_{\text{reduced}} = \left\{ f \in \mathscr{F}, \hat{R}_n(f) = 0 \right\} \subset \mathscr{F}$$

- The generalization puzzle



** Zhou, Sutherland, Srebro. On uniform convergence and low-norm interpolation learning.*

Solution: Capacity Reduced Model Class

What's the **exact** gap between

- Generalization Error
- Uniform Convergence
- Capacity Reduced Uniform Convergence ?

Zhou, Sutherland, Srebro: On uniform convergence and low-norm interpolation learning.

Random Features Model

- Random Features Model: functions $f : \mathbb{R}^d \to \mathbb{R}$

$$\mathscr{F}_{\mathsf{RF}}\,(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N, A) = \left\{ f(\boldsymbol{x}; \boldsymbol{a}) = \sum_{j=1}^{N} a_j \sigma\left(\langle \boldsymbol{x}, \boldsymbol{\theta}_j \rangle\right), \|\boldsymbol{a}\|_2 \leq A, \boldsymbol{a} \in \mathbb{R}^N \right\},$$

where $\boldsymbol{\theta}_j \sim \text{Unif}(\mathbb{S}^{d-1})$ are random weights.

Random Features Model

- Random Features Model: functions $f : \mathbb{R}^d \to \mathbb{R}$

$$\mathscr{F}_{\mathsf{RF}}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N, A) = \left\{ f(\boldsymbol{x}; \boldsymbol{a}) = \sum_{j=1}^{N} a_j \sigma\left(\langle \boldsymbol{x}, \boldsymbol{\theta}_j \rangle\right), \|\boldsymbol{a}\|_2 \le A, \boldsymbol{a} \in \mathbb{R}^N \right\},$$

where $\boldsymbol{\theta}_j \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ are random weights.

- Given a dataset $(\boldsymbol{x}_i, y_i)_{i=1\ldots n}$, where $\boldsymbol{x}_i \overset{i.i.d.}{\sim} \mathrm{Unif}(\mathbb{S}^{d-1})$ and y_i are noisy observations the signal $\boldsymbol{\theta}_\star$: $y_i \sim \mathcal{N}\left(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_\star \rangle, \tau^2\right)$, define

Random Features Model

- Random Features Model: functions $f : \mathbb{R}^d \to \mathbb{R}$

$$\mathscr{F}_{\mathsf{RF}}\left(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N, A\right) = \left\{ f(\boldsymbol{x}; \boldsymbol{a}) = \sum_{j=1}^{N} a_j \sigma\left(\langle \boldsymbol{x}, \boldsymbol{\theta}_j \rangle\right), \|\boldsymbol{a}\|_2 \leq A, \boldsymbol{a} \in \mathbb{R}^N \right\},$$

where $\boldsymbol{\theta}_j \sim \text{Unif}(\mathbb{S}^{d-1})$ are random weights.

- Given a dataset $(\boldsymbol{x}_i, y_i)_{i=1\ldots n}$, where $\boldsymbol{x}_i \overset{i.i.d.}{\sim} \text{Unif}(\mathbb{S}^{d-1})$ and y_i are noisy observations the signal $\boldsymbol{\theta}_\star$: $y_i \sim \mathcal{N}\left(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_\star \rangle, \tau^2\right)$, define

$$\hat{R}_n(\boldsymbol{a}) = \frac{1}{n} \sum_{i=1}^{n} \left[y_i - \sum_{i=1}^{N} a_i \sigma\left(\langle \boldsymbol{x}, \boldsymbol{\theta}_i \rangle\right) \right]^2,$$

Random Features Model

- Random Features Model: functions $f : \mathbb{R}^d \to \mathbb{R}$

$$\mathscr{F}_{\mathsf{RF}}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N, A) = \left\{ f(\boldsymbol{x}; \boldsymbol{a}) = \sum_{j=1}^{N} a_j \sigma\left(\langle \boldsymbol{x}, \boldsymbol{\theta}_j \rangle\right), \|\boldsymbol{a}\|_2 \le A, \boldsymbol{a} \in \mathbb{R}^N \right\},$$

where $\boldsymbol{\theta}_j \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ are random weights.

- Given a dataset $(\boldsymbol{x}_i, y_i)_{i=1\ldots n}$, where $\boldsymbol{x}_i \overset{i.i.d.}{\sim} \mathrm{Unif}(\mathbb{S}^{d-1})$ and y_i are noisy observations the signal $\boldsymbol{\theta}_\star$: $y_i \sim \mathcal{N}\left(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_\star \rangle, \tau^2\right)$, define

$$\hat{R}_n(\boldsymbol{a}) = \frac{1}{n} \sum_{i=1}^{n} \left[y_i - \sum_{i=1}^{N} a_i \sigma\left(\langle \boldsymbol{x}, \boldsymbol{\theta}_i \rangle\right) \right]^2, \qquad R(\boldsymbol{a}) = \mathbb{E}_{\boldsymbol{x},y} \left[y_i - \sum_{i=1}^{N} a_i \sigma\left(\langle \boldsymbol{x}, \boldsymbol{\theta}_i \rangle\right) \right]^2.$$

Random Features Model

- Random Features Model: functions $f : \mathbb{R}^d \to \mathbb{R}$

$$\mathscr{F}_{\mathsf{RF}}\left(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N, A\right) = \left\{ f(\boldsymbol{x}; \boldsymbol{a}) = \sum_{j=1}^{N} a_j \sigma\left(\langle \boldsymbol{x}, \boldsymbol{\theta}_j \rangle\right), \|\boldsymbol{a}\|_2 \le A, \boldsymbol{a} \in \mathbb{R}^N \right\},$$

where $\boldsymbol{\theta}_j \sim \text{Unif}(\mathbb{S}^{d-1})$ are random weights.

- Given a dataset $(\boldsymbol{x}_i, y_i)_{i=1\ldots n}$, where $\boldsymbol{x}_i \overset{i.i.d.}{\sim} \text{Unif}(\mathbb{S}^{d-1})$ and y_i are noisy observations the signal $\boldsymbol{\theta}_\star$: $y_i \sim \mathcal{N}\left(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_\star \rangle, \tau^2\right)$, define

$$\hat{R}_n(\boldsymbol{a}) = \frac{1}{n} \sum_{i=1}^{n} \left[y_i - \sum_{i=1}^{N} a_i \sigma\left(\langle \boldsymbol{x}, \boldsymbol{\theta}_i \rangle\right) \right]^2, \qquad R(\boldsymbol{a}) = \mathbb{E}_{\boldsymbol{x},y} \left[y_i - \sum_{i=1}^{N} a_i \sigma\left(\langle \boldsymbol{x}, \boldsymbol{\theta}_i \rangle\right) \right]^2.$$

- When $N > n$, the min-norm interpolator exists w.h.p.

$$\boldsymbol{a}_{\min} = \arg \min_{\hat{R}_n(\boldsymbol{a})=0} \|\boldsymbol{a}\|.$$

Random Features Model

What's the **exact** gap between

- Generalization Error
- Uniform Convergence
- Capacity Reduced Uniform Convergence

?

Uniform Convergence in Random Features Model

- Generalization Error

$$R(N, n, d) = R\left(\boldsymbol{a}_{\min}\right) - \underbrace{\hat{R}_n(\boldsymbol{a}_{\min})}_{=0}$$

Uniform Convergence in Random Features Model

- Generalization Error

$$R(N, n, d) = R\left(\boldsymbol{a}_{\min}\right) - \underbrace{\hat{R}_n(\boldsymbol{a}_{\min})}_{=0}$$

- Uniform Convergence

$$U(A, N, n, d) = \sup_{(N/d)\|\boldsymbol{a}\|_2^2 \le A} R(\boldsymbol{a}) - \hat{R}_n(\boldsymbol{a})$$

Uniform Convergence in Random Features Model

- Generalization Error

$$R(N, n, d) = R\left(\boldsymbol{a}_{\min}\right) - \underbrace{\hat{R}_n(\boldsymbol{a}_{\min})}_{=0}$$

- Uniform Convergence

$$U(A, N, n, d) = \sup_{(N/d)\|\boldsymbol{a}\|_2^2 \le A} R(\boldsymbol{a}) - \hat{R}_n(\boldsymbol{a})$$

- Uniform Convergence over Interpolators

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

Uniform Convergence in Random Features Model

- Generalization Error

$$R(N, n, d) = R\left(\boldsymbol{a}_{\min}\right) - \underbrace{\hat{R}_n(\boldsymbol{a}_{\min})}_{=0}$$

- Uniform Convergence

$$U(A, N, n, d) = \sup_{(N/d)\|\boldsymbol{a}\|_2^2 \le A} R(\boldsymbol{a}) - \hat{R}_n(\boldsymbol{a})$$

- Uniform Convergence over Interpolators

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

- With the choice of $A = \alpha \cdot (N/d)\|\boldsymbol{a}_{\min}\|_2^2$ (e.g. $\alpha = 1.1$).

Uniform Convergence in Random Features Model

- Generalization Error

$$R(N, n, d) = R\left(\boldsymbol{a}_{\min}\right) - \underbrace{\hat{R}_n(\boldsymbol{a}_{\min})}_{=0}$$

- Uniform Convergence

$$U(A, N, n, d) = \sup_{(N/d)\|\boldsymbol{a}\|_2^2 \le A} R(\boldsymbol{a}) - \hat{R}_n(\boldsymbol{a})$$

- Uniform Convergence over Interpolators

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

- With the choice of $A = \alpha \cdot (N/d)\|\boldsymbol{a}_{\min}\|_2^2$ (e.g. $\alpha = 1.1$).

- Consider $\lim_{d\to\infty} N/d = \psi_1, \quad \lim_{d\to\infty} n/d = \psi_2$.

Uniform Convergence in Random Features Model

- Generalization Error

$$R(N, n, d) = R\left(\boldsymbol{a}_{\min}\right) - \underbrace{\hat{R}_n(\boldsymbol{a}_{\min})}_{=0}$$

- Uniform Convergence

$$U(A, N, n, d) = \sup_{(N/d)\|\boldsymbol{a}\|_2^2 \le A} R(\boldsymbol{a}) - \hat{R}_n(\boldsymbol{a})$$

- Uniform Convergence over Interpolators

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

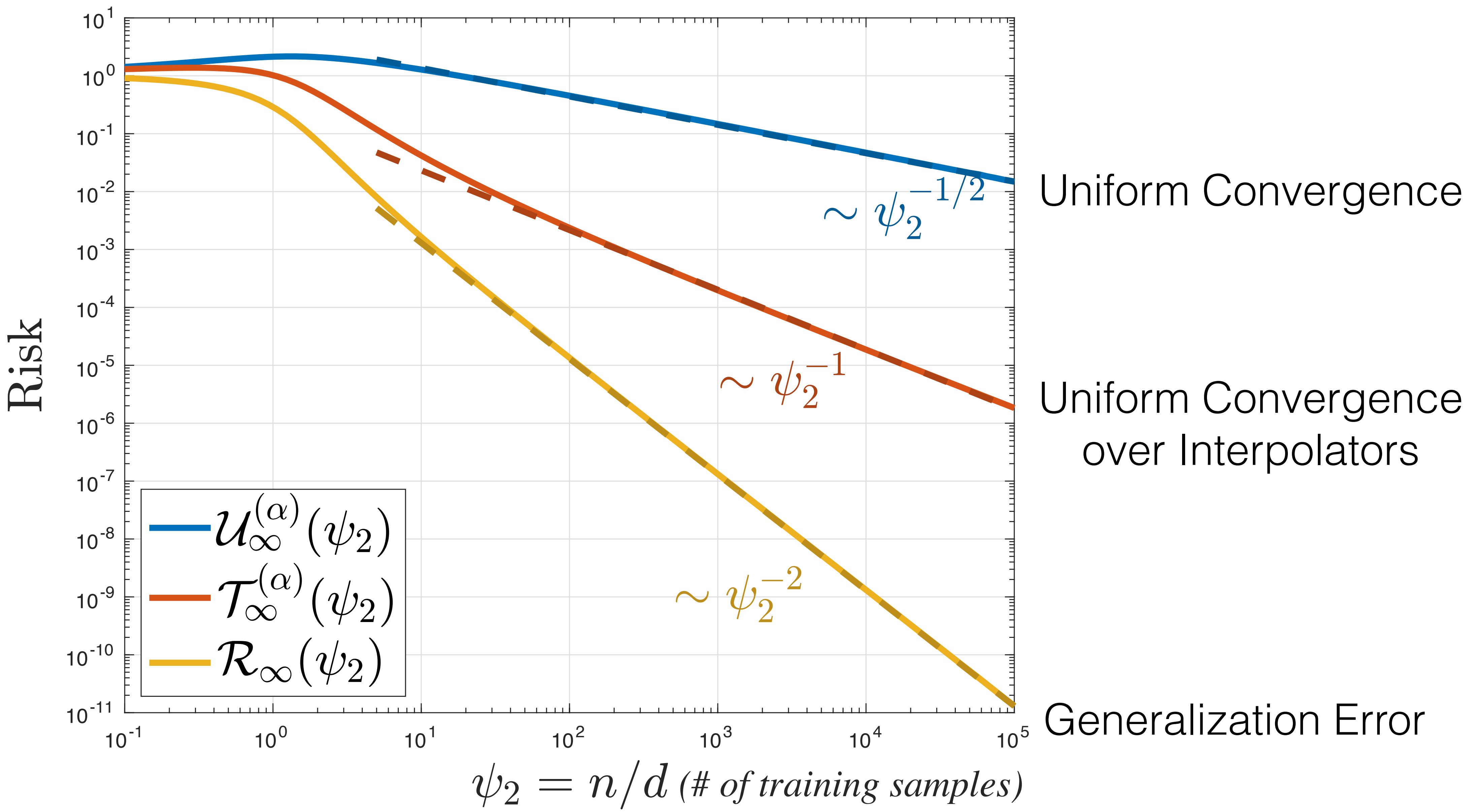
- With the choice of $A = \alpha \cdot (N/d)\|\boldsymbol{a}_{\min}\|_2^2$ (e.g. $\alpha = 1.1$).

- Consider $\lim_{d\to\infty} N/d = \psi_1, \quad \lim_{d\to\infty} n/d = \psi_2$.

  Allows exact computation instead non-asymptotic bounds.

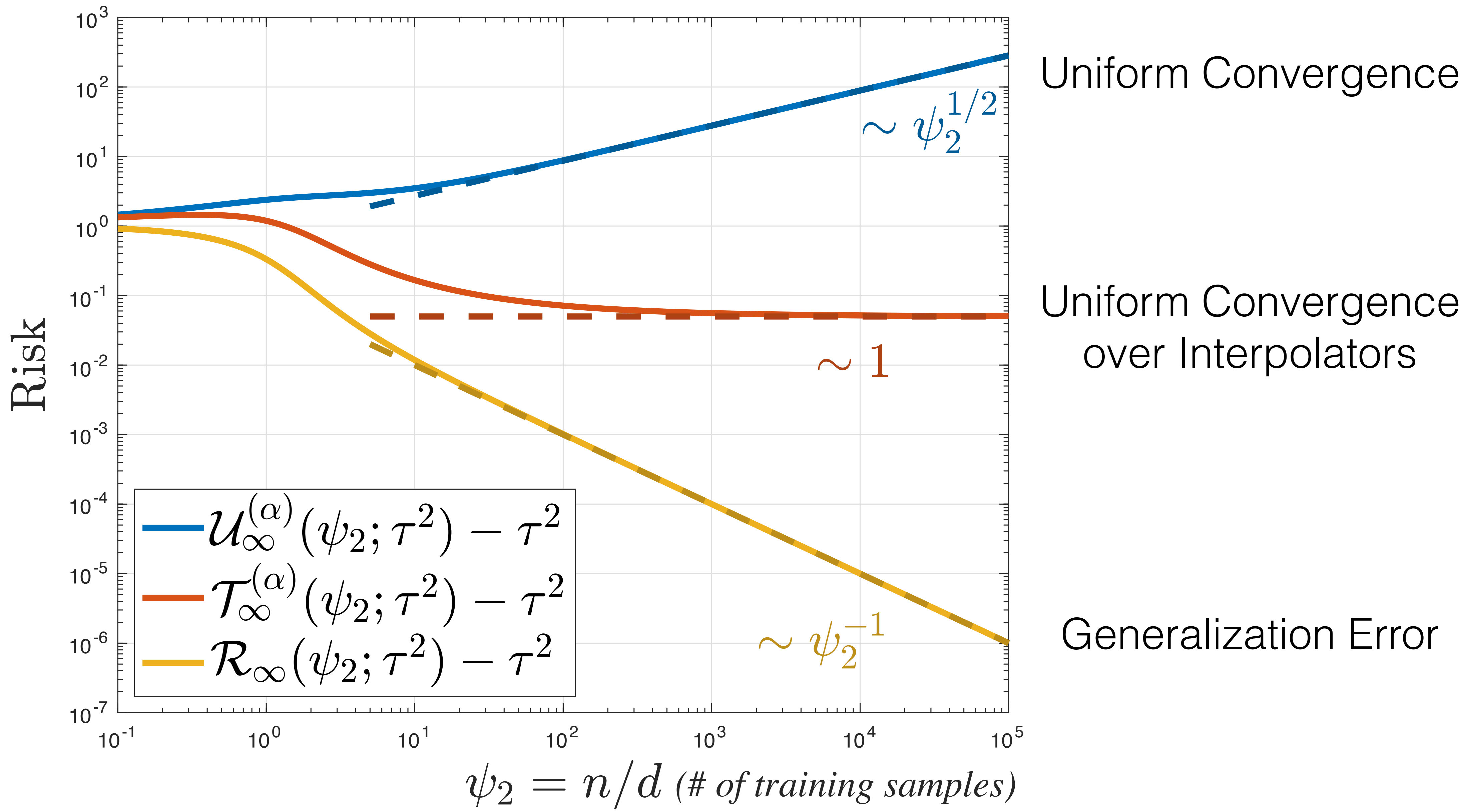
Main Results: Exact Analytical Formulae

- In the noiseless regime $\tau^2 = 0$



Main Results: Exact Analytical Formulae

- In the noisy regime $\tau^2 > 0$



Thanks!