Discussion on
generalization and margins

Daniel Hsu
Columbia University

August 5, 2020
JSM session on “Theoretical Advances In Deep Learning”
What is generalization in machine deep learning?

At least two definitions for generalization error are floated in the community:

1. Out-of-sample (test) error rate $\text{err}(f)$
2. Difference between out-of-sample (test) and in-sample (training) error rates $\text{err}(f) - \text{err}(f; S_n)$

We care about the former, empirical process theory is good for the latter ($S_n \sim \text{Pr}_{x,y}^n$; "uniform convergence bounds")

Major use case: Analysis of Empirical Risk Minimization (ERM) $\min_{f \in F} \text{err}(f; S_n)$
What is generalization in machine deep learning?

At least two definitions for generalization error are floated in the community:

1. Out-of-sample (test) error rate
   \[ \text{err}(f) \]

2. Difference between out-of-sample (test) and in-sample (training) error rates
   \[ \text{err}(f) - \text{err}(f; S_n) \]
What is generalization in machine deep learning?

At least two definitions for generalization error are floated in the community:

1. Out-of-sample (test) error rate
   \[ \text{err}(f) \]

2. Difference between out-of-sample (test) and in-sample (training) error rates
   \[ \text{err}(f) - \text{err}(f; S_n) \]

We care about the former, empirical process theory is good for the latter (since \( S_n \sim (\text{Pr}_{x,y})^n; \) “uniform convergence bounds”)

What is generalization in machine deep learning?

At least two definitions for *generalization error* are floated in the community:

1. Out-of-sample (test) error rate
   \[ \text{err}(f) \]

2. Difference between out-of-sample (test) and in-sample (training) error rates
   \[ \text{err}(f) - \text{err}(f; S_n) \]

We care about the former, empirical process theory is good for the latter (since \( S_n \sim (\Pr_{x,y})^n \); “uniform convergence bounds”)

Major use case: Analysis of Empirical Risk Minimization (ERM)

\[ \min_{f \in \mathcal{F}} \text{err}(f; S_n) \]
Consistency of models that perfectly fit training data

[Belkin, H., Mitra, 2018]: “Weighted & Interpolating $k_n$-NN”
classifier $f_n \equiv f_{S_n}$ satisfies

$$
\mathbb{E}_{S_n} \left[ \Pr_x \left( f_n(x) \neq f_{\text{bayes}}(x) \right) \right] \to 0 \quad \text{as } n \to \infty
$$

under regularity conditions on distribution of $x$
Consistency of models that perfectly fit training data

[Belkin, H., Mitra, 2018]: “Weighted & Interpolating $k_n$-NN”
classifier $f_n \equiv f_{S_n}$ satisfies

$$E_{S_n} \left[ \Pr_x \left( f_n(x) \neq f_{\text{bayes}}(x) \right) \right] \to 0 \quad \text{as } n \to \infty$$

under regularity conditions on distribution of $x$

- In particular,
  $$\text{err}(f_n; S_n) = 0 \quad \text{(always)}$$
  and
  $$E_{S_n} \left[ |\text{err}(f_n) - \text{err}(f_n; S_n)| \right] = E_{S_n} [\text{err}(f_n)] \to \text{err}(f_{\text{bayes}}).$$
Consistency of models that perfectly fit training data

[Belkin, H., Mitra, 2018]: “Weighted & Interpolating $k_n$-NN”
classifier $f_n \equiv f_{S_n}$ satisfies

$$E_{S_n} \left[ \Pr_x \left( f_n(x) \neq f_{\text{bayes}}(x) \right) \right] \to 0 \quad \text{as } n \to \infty$$

under regularity conditions on distribution of $x$

- In particular,

$$\text{err}(f_n; S_n) = 0 \quad \text{(always)}$$

and

$$E_{S_n} \left[ |\text{err}(f_n) - \text{err}(f_n; S_n)| \right] = E_{S_n} [\text{err}(f_n)]$$

$$\to \text{err}(f_{\text{bayes}}).$$

- ∴ Any uniform convergence bound that applies to $f_n$ must
  “stall” at the Bayes error rate (which may be non-zero).
Consistency of models that perfectly fit training data

[Belkin, H., Mitra, 2018]: “Weighted & Interpolating \( k_n \)-NN”
classifier \( f_n \equiv f_{S_n} \) satisfies

\[
\mathbb{E}_{S_n} \left[ \Pr_x \left( f_n(x) \neq f_{\text{bayes}}(x) \right) \right] \to 0 \quad \text{as } n \to \infty
\]

under regularity conditions on distribution of \( x \)

► In particular,

\[
\text{err}(f_n; S_n) = 0 \quad \text{(always)}
\]

and

\[
\mathbb{E}_{S_n} \left[ |\text{err}(f_n) - \text{err}(f_n; S_n)| \right] = \mathbb{E}_{S_n} [\text{err}(f_n)]
\]

\[
\to \text{err}(f_{\text{bayes}}).
\]

► \( \therefore \) Any uniform convergence bound that applies to \( f_n \) must

“stall” at the Bayes error rate (which may be non-zero).

► (Similar results for squared-error regression.)
Uniform convergence and perfect-fit classifiers

Are there issues when $\text{err}(f_{\text{bayes}}) \approx 0$?
Are there issues when $\text{err}(f_{\text{bayes}}) \approx 0$?

**Theisen, Klusowski, Mahoney:** Compelling MNIST analysis

- Many classifiers with $\text{err}(f, S_n) = 0$ have low $\text{err}(f)$
- There are classifiers with $\text{err}(f, S_n) = 0$ and high $\text{err}(f)$
- Therefore, “Uniform convergence bounds” still have problems . . .
Are there issues when $err(f_{\text{bayes}}) \approx 0$?

**Theisen, Klusowski, Mahoney:** Compelling MNIST analysis

- Many classifiers with $err(f, S_n) = 0$ have low $err(f)$
- There are classifiers with $err(f, S_n) = 0$ and high $err(f)$
- $\therefore$ “Uniform convergence bounds” still have problems . . .

Possible fix: Only consider large margin classifiers (or other quantitative inductive bias)

- Schapire, Freund, Bartlett, and Lee (1998); Zhang (2002); . . .
- But *a posteriori* bounds don’t directly analyze the inductive bias achieved by the fitted model
- Sharp contrast with analyses of Ji and Telgarsky; Ji and Telgarsky; Liang, Rakhlin, Zhai; Liang and Sur; . . .
PAC-Bayes approach to margin bounds (e.g., Langford and Shawe-Taylor, 2002) is a relevant bridge between worst-case and average-case analysis.

- Relevance: Maybe practitioners don’t pick a (consistent) classifier at random
BTW

PAC-Bayes approach to margin bounds (e.g., Langford and Shawe-Taylor, 2002) is a relevant bridge between worst-case and average-case analysis.

▶ Relevance: Maybe practitioners don’t pick a (consistent) classifier at random
▶ (But still has same issues as other *a posteriori* bounds.)
Support vector machines (SVMs)

Figure 1: Relevance

I had a joke about SVMs but no one cares any more
Vapnik (1979): mathematical definition of maximum margin linear classifier, along with a theory of generalization.

\[
\min_{w \in \mathbb{R}^d} \|w\|_2 \\
\text{s.t.} \quad y_i x_i^T w \geq 1, \quad i = 1, \ldots, n.
\]

(All \(y_i \in \{-1, 1\}\).)

Figure 1: Relevance
Support vector machines (SVMs)

Vapnik (1979): mathematical definition of maximum margin linear classifier, along with a theory of generalization.

\[
\min_{w \in \mathbb{R}^d} \|w\|_2 \\
\text{s.t. } y_i x_i^T w \geq 1, \quad i = 1, \ldots, n.
\]

(All \(y_i \in \{-1, 1\}\).)

- Why not \(\min_{w \in \mathbb{R}^d} \|w\|_2 \text{ s.t. } x_i^T w = y_i \) (interpolation)?

Figure 1: Relevance
SVMs vs interpolation [Muthukumar et al, 2020]

\[ K(x_1, x_2) = \sum_{k \geq 0} \frac{\sin(kx_1) \sin(kx_2) + \cos(kx_1) \cos(kx_2)}{(k + 1)^2m} \]

Figure 2: SVM solution vs least norm interpolation \((m = 1.5)\)
Margins in very high-dimensions

- Toy setup similar to that of Theisen, Klusowski, Mahoney

\[(x_i, y_i) \sim_{iid} \frac{1}{2}(N_+, 1) + \frac{1}{2}(N_-, -1) \quad i = 1, \ldots, n\]

\[N_+ = \mathcal{N}(\mu, I_d)\]

\[N_- = \mathcal{N}(-\mu, I_d)\]

\[\text{If } d \gg n \log n, \text{ then with high probability, every training example is a support vector:}\]

\[x_i \mathcal{N}_i w_{\text{svm}} = y_i, \quad i = 1, \ldots, n\]

In this case:

- Minimum norm interpolation = SVM solution.

(Similar behavior under anisotropic (subgaussian or subsampled Haar) designs if covariance eigenvalues decay slowly enough)

What about kernels that matter?
Margins in very high-dimensions

- Toy setup similar to that of Theisen, Klusowski, Mahoney

\[(x_i, y_i) \sim_{iid} \frac{1}{2}(N_+, 1) + \frac{1}{2}(N_-, -1) \quad i = 1, \ldots, n\]

\[N_+ = \mathcal{N}(\mu, I_d)\]

\[N_- = \mathcal{N}(-\mu, I_d)\]

- [Muthukumar et al, 2020] and [H., Muthukumar, Xu]:

If \( d \gg n \log n \), then with high probability, every training example is a support vector:

\[x_i^T w_{svm} = y_i, \quad i = 1, \ldots, n\]

where \( w_{svm} \) is the SVM solution.
Margins in very high-dimensions

▶ Toy setup similar to that of Theisen, Klusowski, Mahoney

\[(x_i, y_i) \sim_{iid} \frac{1}{2}(N_+, 1) + \frac{1}{2}(N_-, -1) \quad i = 1, \ldots, n\]

\[N_+ = \mathcal{N}(\mu, I_d)\]

\[N_- = \mathcal{N}(-\mu, I_d)\]

▶ [Muthukumar et al, 2020] and [H., Muthukumar, Xu]: If \(d \gg n \log n\), then with high probability, every training example is a support vector:

\[x_i^T w_{svm} = y_i, \quad i = 1, \ldots, n\]

where \(w_{svm}\) is the SVM solution.

▶ In this case: Minimum norm interpolation = SVM solution.
Margins in very high-dimensions

- Toy setup similar to that of Theisen, Klusowski, Mahoney

\[(x_i, y_i) \sim_{iid} \frac{1}{2}(N_+, 1) + \frac{1}{2}(N_-, -1) \quad i = 1, \ldots, n\]

\[N_+ = \mathcal{N}(\mu, I_d)\]

\[N_- = \mathcal{N}(-\mu, I_d)\]

- [Muthukumar et al, 2020] and [H., Muthukumar, Xu]:
If \(d \gg n \log n\), then with high probability, every training example is a support vector:

\[x_i^\top w_{svm} = y_i, \quad i = 1, \ldots, n\]

where \(w_{svm}\) is the SVM solution.

- In this case: **Minimum norm interpolation = SVM solution.**
- (Similar behavior under anisotropic (subgaussian or subsampled Haar) designs if covariance eigenvalues decay slowly enough)
Margins in very high-dimensions

- Toy setup similar to that of Theisen, Klusowski, Mahoney

\[(x_i, y_i) \sim_{\text{iid}} \frac{1}{2}(N_+, 1) + \frac{1}{2}(N_-, -1) \quad i = 1, \ldots, n\]

\[N_+ = \mathcal{N}(\mu, I_d)\]

\[N_- = \mathcal{N}(-\mu, I_d)\]

- [Muthukumar et al, 2020] and [H., Muthukumar, Xu]:
  If \(d \gg n \log n\), then with high probability, every training example is a support vector:

\[x_i^T w_{\text{svm}} = y_i, \quad i = 1, \ldots, n\]

where \(w_{\text{svm}}\) is the SVM solution.

- In this case: \textit{Minimum norm interpolation = SVM solution.}

- (Similar behavior under anisotropic (subgaussian or subsampled Haar) designs if covariance eigenvalues decay slowly enough)

- What about kernels that matter?
ReLUs in the Neural Tangent Kernel (NTK) regime

**Goal:** Prove error rate $\epsilon \in (0, 1)$ is achieved by training a neural net via gradient descent on an empirical risk.
Goal: Prove error rate $\epsilon \in (0, 1)$ is achieved by training a neural net via gradient descent on an empirical risk

- NTK: kernel-based characterization of certain training processes
ReLU nets in the Neural Tangent Kernel (NTK) regime

**Goal**: Prove error rate $\epsilon \in (0, 1)$ is achieved by training a neural net via gradient descent on an empirical risk

- NTK: kernel-based characterization of certain training processes
- Flurry of results about wide ReLU nets in the NTK regime
ReLU nets in the Neural Tangent Kernel (NTK) regime

**Goal**: Prove error rate $\epsilon \in (0, 1)$ is achieved by training a neural net via gradient descent on an empirical risk

- NTK: kernel-based characterization of certain training processes
- Flurry of results about wide ReLU nets in the NTK regime
  - $\text{poly}(1/\epsilon)$ width is sufficient for *regression* on noise-free data
ReLU nets in the Neural Tangent Kernel (NTK) regime

**Goal**: Prove error rate $\epsilon \in (0, 1)$ is achieved by training a neural net via gradient descent on an empirical risk

- NTK: kernel-based characterization of certain training processes
- Flurry of results about wide ReLU nets in the NTK regime
  - $\text{poly}(1/\epsilon)$ width is sufficient for *regression* on noise-free data
- Ji and Telgarsky:
Goal: Prove error rate $\epsilon \in (0, 1)$ is achieved by training a neural net via gradient descent on an empirical risk

- NTK: kernel-based characterization of certain training processes
- Flurry of results about wide ReLU nets in the NTK regime
  - $\text{poly}(1/\epsilon)$ width is sufficient for regression on noise-free data
- Ji and Telgarsky:
  - $\text{poly log}(1/\epsilon)$ width is sufficient for classification on data separable with a margin
ReLU nets in the Neural Tangent Kernel (NTK) regime

**Goal:** Prove error rate $\epsilon \in (0, 1)$ is achieved by training a neural net via gradient descent on an empirical risk

- **NTK:** kernel-based characterization of certain training processes
- **Flurry of results about wide ReLU nets in the NTK regime**
  - $\text{poly}(1/\epsilon)$ width is sufficient for *regression* on noise-free data
- **Ji and Telgarsky:**
  - $\text{poly log}(1/\epsilon)$ width is sufficient for *classification* on data separable with a margin
  - Margin is defined with respect to infinite-width NTK
ReLU nets in the Neural Tangent Kernel (NTK) regime

**Goal:** Prove error rate $\epsilon \in (0, 1)$ is achieved by training a neural net via gradient descent on an empirical risk

- **NTK:** kernel-based characterization of certain training processes
- Flurry of results about wide ReLU nets in the NTK regime
  - $\text{poly}(1/\epsilon)$ width is sufficient for *regression* on noise-free data
- **Ji and Telgarsky:**
  - $\text{poly log}(1/\epsilon)$ width is sufficient for *classification* on data separable with a margin
  - Margin is defined with respect to infinite-width NTK
- A pressing question: Is gap the real?
**Goal**: Prove error rate $\epsilon \in (0, 1)$ is achieved by training a neural net via gradient descent on an empirical risk

- **NTK**: kernel-based characterization of certain training processes
- Flurry of results about wide ReLU nets in the NTK regime
  - $\text{poly}(1/\epsilon)$ width is sufficient for *regression* on noise-free data
- **Ji and Telgarsky**:
  - $\text{poly log}(1/\epsilon)$ width is sufficient for *classification* on data separable with a margin
  - Margin is defined with respect to infinite-width NTK
- A pressing question: Is gap the real?
  - (Telgarsky: “It’s subtle . . .”)

8 / 9
Parting words

Uniform convergence + ERM go hand-in-hand
Parting words

Uniform convergence + ERM go hand-in-hand

- Uniform convergence makes sense ⇔ empirical risk makes sense
- If empirical risk = 0 always, maybe look elsewhere

Thank you!

Thanks to: NSF CCF-1740833, DMR-153491; Sloan Research Fellowship; Simons Institute for the Theory of Computing Sp'17 & Su'19 programs
Parting words

Uniform convergence + ERM go hand-in-hand

▶ Uniform convergence makes sense ⇔ empirical risk makes sense
▶ If empirical risk = 0 always, maybe look elsewhere

Shift focus of analysis to inductive bias (e.g., margins)
Parting words

Uniform convergence + ERM go hand-in-hand

- Uniform convergence makes sense $\iff$ empirical risk makes sense
- If empirical risk $= 0$ always, maybe look elsewhere

Shift focus of analysis to inductive bias (e.g., margins)

Thank you!

Thanks to: NSF CCF-1740833, DMR-153491; Sloan Research Fellowship; Simons Institute for the Theory of Computing Sp’17 & Su’19 programs