In Defense of Uniform Convergence: Generalization via derandomization with an application to interpolating predictors (Jeffrey Negrea, 2020 ICML)

Presented by Insung Kong

2022 1/17

Seoul National University

Table of Contents

1. Introduction
2. Preliminaries
3. Glivenko–Cantelli property and surrogate
4. Example : overparameterized linear regression

Introduction

Introduction

- One of the central problems in learning theory is to explain the statistical performance of deep learning algorithms.
- The bulk of recent work on this problem implicitly assumes the classifier learned by SGD belongs to a class for which there is a uniform and tight bound on the generalization error.
- The traditional notions of uniform convergence (Glivenko–Cantelli classes) are may not applicable.

Introduction

- Paper extend the concept of uniform convergence to the setting of sequences of learning problems of increasing complexity, defined by the structural Glivenko–Cantelli property.
- They introduce a approach to relate sequences of learning problems which are not structural Glivenko–Cantelli to ones that may be.
- Paper applicated it to overparameterized linear regression.

Preliminaries

Notations

- $Z_1, \ldots, Z_n$: i.i.d. random elements in a space $\mathcal{S}$ with common distribution $\mathcal{D}$
- $S = (Z_1, \ldots, Z_n)$: the training sample
- Loss function $l : \mathcal{H} \times \mathcal{S} \to \mathbb{R}_+$ for a space $\mathcal{H}$ of hypotheses.
- $\mathcal{M}_1(\mathcal{H})$: space of distribution on $\mathcal{H}$.
- For $Q \in \mathcal{M}_1(\mathcal{H})$, the loss and risk are defined to be

$$\ell(Q, z) = \int \ell(h, z) Q(\mathrm{d}h), \quad L_{\mathcal{D}}(Q) = \int \ell(Q, z) \mathcal{D}(\mathrm{d}z)$$

- $L_S(Q) = L_{\widehat{\mathcal{D}}_n}(Q)$: empirical risk
- For $h \in \mathcal{H}$ $L_{\mathcal{D}}(h) = L_{\mathcal{D}}(\delta_h)$ and $L_{\mathcal{S}}(h) = L_{\mathcal{S}}(\delta_h)$
- $\hat{h} \in \mathcal{H}$: learned classifier, which is random.

Glivenko–Cantelli property and surrogate

Glivenko–Cantelli property

Definition 3.1
Let $(\mathcal{S}, \mathcal{F}, \mathcal{D})$ be a probability space. Let $\mathcal{H}$ be a collection of measurable functions on $(\mathcal{S}, \mathcal{F}, \mathcal{D})$.

Then $\mathcal{H}$ has the Glivenko–Cantelli property if

$$\lim_{p \to \infty} \mathbb{E}\left[\sup_{h \in \mathcal{H}} \left|\mathcal{D}h - \widehat{\mathcal{D}}_p h\right|\right] = 0$$

where $Ph = \int h(x)P(dx)$ and $\widehat{\mathcal{D}}_p$ is the empirical distribution of an IID sample of size p from $\mathcal{D}$.

Definition 3.2
Let $\left\{\left(\mathcal{S}^{(p)}, \mathcal{F}^{(p)}, \mathcal{D}^{(p)}\right)\right\}_{p\in\mathbb{N}}$ be a sequence of probability space. Let $\mathcal{H}^{(p)}$ be a collection of measurable functions on $\left(\mathcal{S}^{(p)}, \mathcal{F}^{(p)}, \mathcal{D}^{(p)}\right)$.

Then $\mathcal{H}^{(\cdot)}$ has the structural $(\mathcal{D}^{(\cdot)}, n_{(\cdot)})$-Glivenko–Cantelli property, if

$$\lim_{p\to\infty} \mathbb{E}\left[\sup_{h\in\mathcal{H}^{(p)}} \left|\mathcal{D}^{(p)}h - \widehat{\mathcal{D}^{(p)}}_{n_p} h\right|\right] = 0$$

where $Ph = \int h(x)P(dx)$ and $\widehat{\mathcal{D}^{(p)}}_p$ is the empirical distribution of an IID sample of size p from $\mathcal{D}^{(p)}$.

Surrogate decomposition

Lemma 3.3 (Surrogate decomposition)
For every random element $\mathcal{Q}$ in $\mathcal{M}_1(\mathcal{H})$

$$\begin{aligned}\mathbb{E}\left[L_{\mathcal{D}}(\hat{h}) - L_S(\hat{h})\right] &= \mathbb{E}\left[L_{\mathcal{D}}(\hat{h}) - L_{\mathcal{D}}(Q)\right] \\ &\quad + \mathbb{E}\left[L_{\mathcal{D}}(Q) - L_S(Q)\right] \\ &\quad + \mathbb{E}\left[L_S(Q) - L_S(\hat{h})\right],\end{aligned}$$

provided the three expectations on the r.h.s. are finite

Example : overparameterized linear regression

Notation

- $X_i \overset{\text{iid}}{\sim} \mathrm{N}_{1\times d}(0, \Sigma_n)$: random row vectors
- $X^\top = (X_1^\top, \ldots, X_n^\top)^\top]$
- $(Y_i \mid X) \overset{\text{ind}}{\sim} \mathrm{N}\left(X_i\beta_n, \sigma^2\right)$.
- $\ell(\beta, (x, y)) = (x\beta - y)^2$.
- $\hat{\beta}(X, Y) = \left(X^\top X\right)^+ X^\top Y$: minimum norm interpolating linear predictor

Lemma 4.1 (Failure of uniform convergence for overparameterized linear regression)

There is no sequence of measurable set $\{A_n\}_{n \in \mathcal{N}}$ *such that* $\mathbb{P}\left((X, Y) \in A_n\right) > 2/3$ *for all* $n \in \mathcal{N}$ *and for which*

$$\limsup_{n \to \infty} \mathbb{E} \sup_{(\tilde{X}, \bar{Y}) \in A_n} \left| L_D(\hat{\beta}(\tilde{X}, \tilde{Y})) - L_S(\hat{\beta}(\tilde{X}, \tilde{Y})) \right| \leq \frac{3}{2} L_D(\beta)$$

We will consider the surrogate given by the minimum norm interpolating predictor for the training data with label noise removed. i.e.

$$\hat{\beta}_0 = \left(X'X\right)^{+} X'X\beta$$

Lemma 4.2 ((The sequence of surrogate hypothesis classes is SGC)

$\left\{\hat{\beta}_0(S) : S \in \mathcal{S}^{(n)}\right\}_{n\in\mathbb{N}}$ *is* $\left(\mathcal{D}^{(n)}, n\right) - SGC$*. Quantitatively, for a universal constant* $C > 0$,

$$\mathbb{E} \sup_{(X_0, Y_0)\in\mathbb{R}^{n\times d}\times\mathbb{R}^n} \left| L_D\left(\hat{\beta}_0\left(X_0, Y_0\right)\right) - L_S\left(\hat{\beta}_0\left(X_0, Y_0\right)\right)\right| \leq C\frac{\sigma^2 + \left\|\beta_n\right\|^2 \left\|\Sigma_n\right\| \max\left(\sqrt{r_0\left(\Sigma_n\right)}, r_0\left(\Sigma_n\right)/\sqrt{n}\right)}{\sqrt{n}}$$

Surrogate decomposition of $\hat{\beta}$

Lemma 4.3 (Surrogate decomposition of $\hat{\beta}$)

$$
\begin{aligned}
&L_D(\hat{\beta}) - L_S(\hat{\beta}) \\
&= \left(L_S\left(\hat{\beta}_0\right) - L_S(\hat{\beta})\right) + \left(L_D(\hat{\beta}) - L_D\left(\hat{\beta}_0\right)\right) \\
&\quad + \left(L_D\left(\hat{\beta}_0\right) - L_S\left(\hat{\beta}_0\right)\right),
\end{aligned}
$$

with

$$
\begin{aligned}
L_S\left(\hat{\beta}_0\right) - L_S(\hat{\beta}) &= \frac{1}{n}\|Z\|^2 \\
L_D(\hat{\beta}) - L_D\left(\hat{\beta}_0\right) &= \operatorname{Tr}\left(X\left(X'X\right)^{+}\Sigma\left(X'X\right)^{+}X'ZZ'\right), \\
L_D\left(\hat{\beta}_0\right) - L_S\left(\hat{\beta}_0\right) &= \sigma^2 - \frac{\|Z\|^2}{n} + \beta' P(X)^{\perp}\Sigma P(X)^{\perp}\beta.
\end{aligned}
$$

Theorem 4.4 (Expected risk bound for overparameterized linear regression)

For a universal constant $C > 0$,

$$\mathbb{E}L_D(\hat{\beta}) \leq \sigma^2 + C\frac{\sigma^2 + \|\beta_n\|^2 \|\Sigma_n\| \max\left(\sqrt{r_0(\Sigma_n)}, r_0(\Sigma_n)/\sqrt{n}\right)}{\sqrt{n}} + c\sigma^2\left(\frac{k_n^*}{n} + \frac{n}{R_{k_n^*}(\Sigma_n)}\right)$$

In particular, if $\{\Sigma_n\}_{n\in\mathbb{N}}$ and $\left\{\|\beta_n\|^2 \|\Sigma_n\|\right\}_{n\in\mathbb{N}}$ is bounded then $\mathbb{E}L_D(\hat{\beta}) \to \sigma^2$.