Theoretical Principles of Deep Learning

Class 4: Generalization

Hédi Hadiji

Université Paris-Saclay - CentraleSupelec

hedi.hadiji@l2s.centralesupelec.fr

January 8th 2024

Table of Contents

Plan

Last time: Optimization

- Some neural nets are easy to optimize in the **lazy regime**
- E.g. very wide nets, or nets with scaled outputs.

Today: Generalization theory. Why/when should good training performance imply good test performance.

Reading Material:

- Telgarsky notes
- Understanding Machine Learning, theory and algorithms

Table of Contents

Setting: Supervised learning

Sample $S = (x_i, y_i)_{i \in [n]}$, i.i.d. from unknown distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$.

Objective of Supervised Learning

Given a sample S, find a hypothesis $h_S : \mathcal{X} \to \mathcal{Y}$ such that the risk

$$R(h_S) := \mathbb{E}_{(X,Y)\sim\mathcal{D}}\left[\ell(h_S(X), Y)\right]$$

is small with high probability.

A standard method is to compute an (approximate) ERM.

ERM

Fix a class of hypotheses $\mathcal{H}$. Look for $h_S \in \mathcal{H}$ with small empirical risk:

$$R_S(h_S) := \frac{1}{n}\sum_{i=1}^{n} \ell(h_S(X_i), Y_i)\,.$$

Generalization gap

Today: we forget about optimization (how we compute ERM) and focus on statistical learning (**"Is ERM any good in terms of true risk?"**)

Definition (Generalization gap)

For a hypothesis $h \in \mathcal{H}$, the generalization gap is the difference between the true risk and the empirical risk on the sample

$$R(h) - R_S(h) .$$

a.k.a. difference between train loss and true loss.

Rest of this class

A generalization bound is an upper bound on the generalization gap of the output of a training algorithm.

Goal of today

Build some general machinery to prove generalization bounds and discuss them when applied to neural networks.

Table of Contents

Generalization for a single hypothesis

Fix a single hypothesis h,

$$R(h) - R_S(h) = \mathbb{E}\left[\ell(h(X), Y)\right] - \frac{1}{n}\sum_{i=1}^{n}\ell(h(X_i), Y_i)$$

For large S, by the central limit theorem, the generalization gap of a single hypothesis is approximately Gaussian with mean 0 and variance C/n.

Generalization for a single hypothesis II

Theorem (Hoeffding's inequality)

If X_i are i.i.d. r.v. bounded in $[0,1]$, then with probability at least $1-\delta$,

$$\mathbb{E}[X] - \bar{X}_n \leqslant \sqrt{\frac{\log(1/\delta)}{2n}} .$$

Assume the loss is bounded in $[0,1]$ then with probability at least $1-\delta$:

$$R(h) \leqslant R_S(h) + \sqrt{\frac{\log(1/\delta)}{2n}} .$$

What if h is the ERM hypothesis on the sample S?

Uniform convergence

One way to prove bounds for the ERM is **uniform convergence**. If

$$\sup_{h\in\mathcal{H}} \big(R(h) - R_S(h)\big) \leqslant B$$

then for any hypothesis $h^\star \in \mathcal{H}$

$$R(h_{ERM}) \leqslant R_S(h_{ERM}) + B \leqslant R_S(h^\star) + B = R(h^\star) + B + (R_S(h^\star) - R(h^\star))$$

$h^\star$ is a single hypothesis, so the final term can be bounded with Hoeffding.

If B is small enough, with enough data points, the ERM can learn as well as the best hypothesis in $\mathcal{H}$.

Generalization in Finite Classes

Theorem (Finite classes)

Fix a sample distribution $\mathcal{D}$ and loss bounded in $[0, 1]$. For any sample distribution, with probability at least $1 - \delta$,

$$\sup_{h \in \mathcal{H}} \left(R(h) - R_S(h) \right) \leqslant \sqrt{\frac{\log(|\mathcal{H}|/\delta)}{2n}}$$

(From now on, “For any sample distribution” implicitly assumed.)

Proof: Board.

Some observations from finite classes

$1/\sqrt{n}$ is the standard dependence on the number of data points. Can also get *fast rates* of order $1/n$ in nicer cases (e.g. low variance labels).

For finite classes, $n \geqslant \log|\mathcal{H}|$ are sufficient for the ERM to start learning.

Bigger classes mean worse bounds: it takes more data points to start having guarantees of learning. But this bound ignores the possible structure of $\mathcal{H}$.

What about infinite classes?

- For binary classification and 0-1 loss, **VC dimension** characterizes the learnability.
- Another approach is to discretize the hypothesis space and compute **covering numbers**.
- We will talk of a more general tool: **Rademacher complexity**.

Table of Contents

Rademacher complexity

One of the central tools to derive generalization in modern theory is

Definition (Rademacher complexity)

Let $(\sigma_i)_{i\in[n]}$ be Rademacher rv (± 1 with prob. $1/2$). The conditional Rademacher complexity on sample S with loss ℓ is

$$\mathcal{R}_S(\ell, \mathcal{H}) = \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{n} \sigma_i \ell(h(X_i), Y_i)\right]$$

and the Rademacher complexity of $\mathcal{H}$ with loss ℓ is

$$\mathcal{R}(\ell, \mathcal{H}) = \mathbb{E}\left[\mathcal{R}_S(\ell, \mathcal{H})\right].$$

Elementary properties

For any sample S,

- $\mathcal{R}_S(\ell, \mathcal{H}) \geqslant 0$
- If $\mathcal{H} = \{h\}$, then $\mathcal{R}_S(\ell, \mathcal{H}) = 0$
- If loss is bounded in $[0, 1]$, then $\mathcal{R}_S(\ell, \mathcal{H}) \leqslant 1$
- If $\mathcal{H}_1 \subset \mathcal{H}_2$, then $\mathcal{R}_S(\ell, \mathcal{H}_1) \leqslant \mathcal{R}_S(\ell, \mathcal{H}_2)$.

In binary classification, if ℓ is the 0-1 loss, then ... Rademacher complexity measures the capacity of the hypothesis class to classify arbitrarily the features.

Generalization bounds

Theorem (Rademacher Generalization Bounds)

With probability at least $1-\delta$*,*

$$R(h) - R_S(h) \leqslant 2\mathcal{R}(\ell, \mathcal{H}) + \sqrt{\frac{\log 1/\delta}{2n}}$$

and with probability at least $1-\delta$*,*

$$R(h) - R_S(h) \leqslant 2\mathcal{R}_S(\ell, \mathcal{H}) + \sqrt{\frac{2\log(2/\delta)}{n}}\,.$$

Proof: Symmetrization + concentration of suprema of empirical processes.

Now it suffices to upper bound the Rademacher complexity of $\mathcal{H}$ to obtain generalization guarantees.

Tool: McDiarmid's inequality

Theorem (McDiarmid's inequality)

Let f be a real-valued function of n points such that for any $z_1, \ldots, z_n$, for any $i \in [n]$ and z_i', we have

$$|f(z_1, \ldots, z_i, \ldots, z_n) - f(z_1, \ldots, z_i', \ldots, z_n)| \leqslant c_i$$

then with probability at least $1 - \delta$,

$$f(Z) - \mathbb{E}[f(Z)] \leqslant \sqrt{\frac{1}{2} \sum_{i=1}^{n} c_i^2 \log(1/\delta)}\,.$$

Rademacher calculus

Rademacher complexity is nice because of tools to upper bound it.
Let V be a set of vectors in $\mathbb{R}^n$, the Rademacher complexity of V is

$$\mathcal{R}(V) = \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{v\in V}\sum_{i=1}^{n}\sigma_i v_i\right]$$

Three main tools for Rademacher manipulations

- Massart’s lemma
- Contraction lemma
- Convex hull lemma

Rademacher toolbox

Proposition (Massart's lemma)

If $|V| \leqslant K$*, then* $\mathcal{R}(V) \leqslant \max_{v \in V} \|v - \bar{v}\| \sqrt{2 \ln K}/n$*, where* $\bar{v}$ *is the average* v*.*

Proposition (Contraction lemma)

Let $\Phi_i : \mathbb{R} \to \mathbb{R}$ *be L-Lipschitz functions, and*
$\Phi : (v_1, \ldots, v_n) \to (\Phi_1(v_1), \ldots, \Phi_n(v_n))$*, then* $\mathcal{R}(\Phi(V)) \leqslant L\mathcal{R}(V)$

Proposition (Convex hulls)

If V *is compact then* $\mathcal{R}(\mathrm{Conv}(V)) = \mathcal{R}(V)$.

Application: Rederivation for finite classes

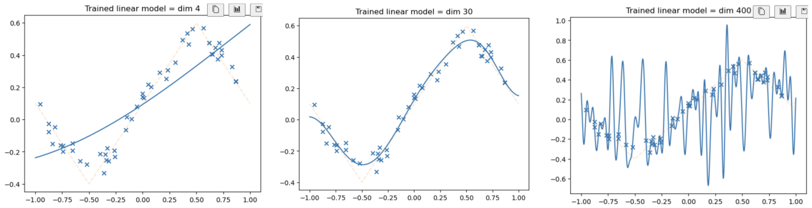
Board.

Table of Contents

Bias/complexity trade-off and overfitting

Typical generalization bounds look like: with probability .99,

$$R(\widehat{h}) \leqslant \underbrace{\inf_{h\in\mathcal{H}} R(h)}_{\searrow \text{ with } \mathcal{H}} + \underbrace{c\sqrt{\frac{\mathrm{Comp}(\mathcal{H})}{n}}}_{\nearrow \text{ with } \mathcal{H}}$$

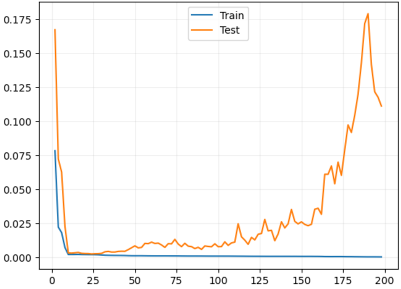


Standard intuition: if $\mathcal{H}$ is not expressive enough, then unable to catch the data. **Underfitting**.

If $\mathcal{H}$ is very expressive, then many ways to fit the data, but might not choose the correct one. The ERM may start fitting noise. **“ Overfitting ”**.

Bias-complexity tradeoff

Trade-off is sometimes real: e.g. least-squares linear regression



Train and test losses vs. dimension of space of regression functions

Not what happens in practice with deep nets. Often there is **double descent**, i.e., performance gets better with more complex models. Bounds that only depend on the number of parameters fail to account for that.

Other approaches

We discussed a type of generalization bound that builds on measuring the model complexity

- Regularization. Training favor 'simple' hypotheses: dropout, layer normalization, data augmentation
- Implicit regularization due to SGD. e.g., we saw last week that in the lazy regime, SGD stays close to initialization.
- Stability analysis: cf. the Perceptron. If an algorithm is not too sensitive to individual data points it should generalize.

Beyond uniform convergence

- PAC-Bayes bounds

Table of Contents

Conclusion and Next time

Today

- Defined generalization
- Introduced a powerful method to derive generalization bounds for many learning settings: Rademacher complexity

In problem session: will apply these to obtain bounds for neural nets.

Next time: **Neural Tangent Kernel**