Generalization in Deep Networks. I.

Peter Bartlett

CS and Statistics
UC Berkeley

19 March, 2019

Statistical Learning Theory

Probabilistic Formulations of Prediction Problems

Aim: Predict an outcome y from some set $\mathcal{Y}$ of possible outcomes, on the basis of some observation x from a feature space $\mathcal{X}$.

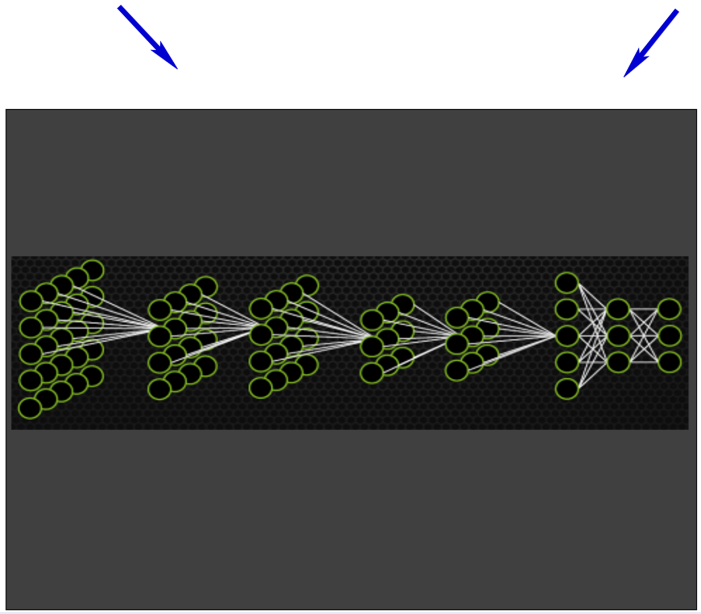
Use *data set* of n pairs:

$$(x_1, y_1), \ldots, (x_n, y_n),$$

to choose a function $f : \mathcal{X} \to \mathcal{Y}$ so that, for subsequent (x, y) pairs, $f(x)$ is a good prediction of y.

Prediction with Deep Networks

Use training data to choose parameters



Probabilistic Formulations of Prediction Problems

To define the notion of a 'good prediction,' we can define a **loss function**

$$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}.$$

$\ell(\hat{y}, y)$ is cost of predicting $\hat{y}$ when the outcome is y.
Aim: $\ell(f(x), y)$ small.

Probabilistic Formulations of Prediction Problems

Example

In *pattern classification* problems, the aim is to classify a pattern x into one of a finite number of classes (that is, the label space $\mathcal{Y}$ is finite). If all mistakes are equally bad, we could define

$$\ell(\hat{y}, y) = 1[\hat{y} \neq y] = \begin{cases} 1 & \text{if } \hat{y} \neq y, \\ 0 & \text{otherwise.} \end{cases}$$

Example

In a *regression* problem, with $\mathcal{Y} = \mathbb{R}$, we might choose the quadratic loss function, $\ell(\hat{y}, y) = (\hat{y} - y)^2$.

Probabilistic Assumptions

Assume:

- There is a probability distribution P on $\mathcal{X} \times \mathcal{Y}$,
- The pairs $(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)$ are chosen independently according to P

The aim is to choose f with small *risk*:

$$R(f) = \mathbb{E}\ell(f(X), Y).$$

For instance, in the pattern classification example, this is the misclassification probability.

$$R(f) = \mathbb{E}1[f(X) \neq Y] = \Pr(f(X) \neq Y).$$

Probabilistic Assumptions

Some things to notice:

1. The distribution P can be viewed as modelling both the relative frequency of different features or covariates X, together with the conditional distribution of the outcome Y given X.
2. The assumption that the data is i.i.d. is a strong one.
   But we need to assume something about what the information in the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ tells us about (X, Y).

Probabilistic Assumptions

3. The function $x \mapsto f_n(x) = f_n(x; X_1, Y_1, \ldots, X_n, Y_n)$ is random, since it depends on the random data $D_n = (X_1, Y_1, \ldots, X_n, Y_n)$. Thus, the risk

$$\begin{aligned} R(f_n) &= \mathbb{E}\left[\ell(f_n(X), Y) | D_n\right] \\ &= \mathbb{E}\left[\ell(f_n(X; X_1, Y_1, \ldots, X_n, Y_n), Y) | D_n\right] \end{aligned}$$

is a random variable. We might aim for $\mathbb{E}R(f_n)$ small, or $R(f_n)$ small with high probability (over the training data).

Key Questions

We might choose f_n from some class F of functions
(for instance, linear function, sparse linear function, ReLU network with fixed architecture and arbitrary parameters, ReLU network with fixed depth and a bound on norms of parameter matrices in each layer, ...).

1. Can we design algorithms for which f_n is close to the best that we could hope for, given that it was chosen from F? (that is, is $R(f_n) - \inf_{f \in F} R(f)$ small?)
2. How does the performance of f_n depend on n? On the complexity of F? On P?
3. Can we ensure that $R(f_n)$ approaches the best possible performance (that is, the infimum over all f of $R(f)$)?

Statistical Learning Theory

- We are concerned with results that apply to large classes of distributions P, such as the set of *all* joint distributions on $\mathcal{X} \times \mathcal{Y}$.
- Typically, we will not assume that P comes from a small (e.g., finite-dimensional) space, $P \in \{P_\theta : \theta \in \Theta\}$.
- We will mostly be concerned with ensuring that the performance is close to the best we can achieve using prediction rules from some fixed class F.

Statistical Learning Theory: Key Issues

Approximation How good is the best f in the class F that we are using? That is, how close to $\inf_f R(f)$ is $\inf_{f \in F} R(f)$?

Estimation How close is our performance to that of the best f in F? (Recall that we only have access to the distribution P through observing a finite data set.)

Computation We need to use the data to choose f_n, typically by solving some kind of optimization problem. How can we do that efficiently?

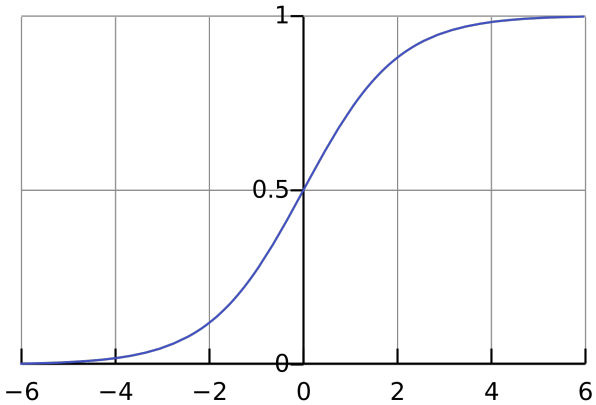
Deep Networks

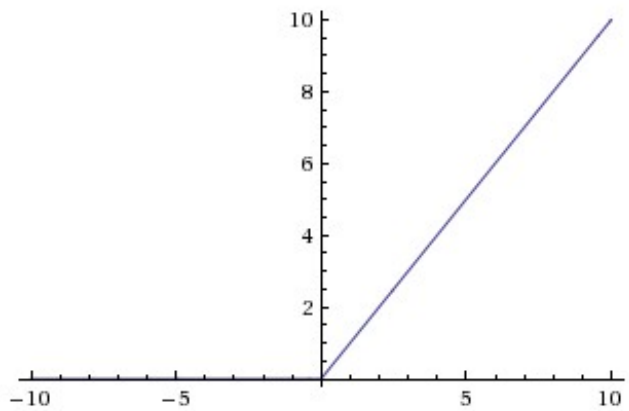
Deep compositions of nonlinear functions

$$h = h_m \circ h_{m-1} \circ \cdots \circ h_1$$

e.g., $h_i : x \mapsto \sigma(W_i x)$ $\qquad$ $h_i : x \mapsto r(W_i x)$

$$\sigma(v)_i = \frac{1}{1 + \exp(-v_i)}, \qquad r(v)_i = \max\{0, v_i\}$$





Why Deep Networks? Some Intuition

Representation learning

Depth provides an effective way of representing useful features.

Rich non-parametric family

Depth provides parsimonious representions.
Nonlinear parameterizations provide better rates of approximation.
Some functions require much more complexity for a shallow representation.

But...

- Optimization?
  - Nonlinear parameterization.
  - Apparently worse as the depth increases.
- Generalization?
  - What determines the statistical complexity of a deep network?

Statistical Learning Theory: Key Issues

- This lecture and the next will focus on the *estimation* issue.
- The third lecture will focus on *computation* issues for deep residual networks.

Outline

- Uniform laws of large numbers
- Rademacher complexity and uniform laws
  (Concentration. Symmetrization. Restrictions.)
- Controlling Rademacher complexity:
  - Growth function
  - VC-dimension
  - Structural results for Rademacher complexity
- Neural networks
  - VC-dimension
  - Large margin classifiers
  - Rademacher averages for sigmoid networks
  - Rademacher averages for ReLU networks
- Interpolating prediction rules

Uniform Laws of Large Numbers: Motivation

Consider the performance of empirical risk minimization:
Choose $f_n \in F$ to minimize $\hat{R}(f)$, where $\hat{R}$ is the *empirical risk*,

$$\hat{R}(f) = P_n \ell(f(X), Y) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i).$$

For pattern classification, this is the proportion of training examples misclassified.
Define $f^* = \arg\min_{f \in F} R(f)$. How does the excess risk, $R(f_n) - R(f^*)$ behave?
We can write

$$R(f_n) - R(f^*) = \left[R(f_n) - \hat{R}(f_n)\right] + \left[\hat{R}(f_n) - \hat{R}(f^*)\right] + \left[\hat{R}(f^*) - R(f^*)\right]$$

Uniform Laws of Large Numbers: Motivation

One of these terms is a difference between a sample average and an expectation for the fixed function $(x, y) \mapsto \ell(f^*(x), y)$:

$$\hat{R}(f^*) - R(f^*) = (P_n - P)\ell(f^*(X), Y)$$

The law of large numbers shows that this term converges to zero; and with information about the tails of $\ell(f^*(X), Y)$ (such as boundedness), we can get bounds on its value.
Another term, $\hat{R}(f_n) - \hat{R}(f^*)$, is non-positive, because f_n is chosen to minimize $\hat{R}$.

Uniform Laws of Large Numbers: Motivation

The other term, $R(f_n) - \hat{R}(f_n)$, is more interesting. For any fixed f, this difference goes to zero. But f_n is random, since it is chosen using the data. An easy upper bound is

$$R(f_n) - \hat{R}(f_n) \leq \sup_{f \in F} \left| R(f) - \hat{R}(f) \right|,$$

and this motivates the study of uniform laws of large numbers.

Uniform Laws of Large Numbers

For a class F of functions $f : \mathcal{X} \to [0, 1]$, suppose that $X_1, \ldots, X_n, X$ are i.i.d. on $\mathcal{X}$, and consider

$$Z = \sup_{f \in F} \left| \mathbb{E} f(X) - \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right| =: \| \underbrace{P - P_n}_{\text{emp proc}} \|_F .$$

If Z converges to 0, this is called a *uniform law of large numbers*.

Glivenko-Cantelli Classes

Definition

F is a **Glivenko-Cantelli class** for P if $\sup_{f \in F} |P_n f - Pf| =: \|P_n - P\|_F \xrightarrow{P} 0$.

- P is a distribution on $\mathcal{X}$,
- $X_1, \ldots, X_n$ are drawn i.i.d. from P,
- P_n is the empirical distribution (which assigns mass $1/n$ to each of $X_1, \ldots, X_n$),
- F is a set of measurable real-valued functions on $\mathcal{X}$ with finite expectation under P,
- $P_n - P$ is an **empirical process**, that is, a stochastic process indexed by a class of functions F, and
- $\|P_n - P\|_F := \sup_{f \in F} |P_n f - Pf|$.

Glivenko-Cantelli Classes

Why 'Glivenko-Cantelli'? An example of a uniform law of large numbers.

Glivenko-Cantelli Theorem

$\|F_n - F\|_\infty \xrightarrow{as} 0.$

Here, F is a cumulative distribution function, F_n is the empirical cumulative distribution function,

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1[X_i \geq x],$$

where $X_1, \ldots, X_n$ are i.i.d. with distribution F, and $\|F - G\|_\infty = \sup_t |F(t) - G(t)|$.

Glivenko-Cantelli Theorem

$\|P_n - P\|_G \xrightarrow{as} 0$, for $G = \{x \mapsto 1[x \leq \theta] : \theta \in \mathbb{R}\}$.

Glivenko-Cantelli Classes

Not all F are Glivenko-Cantelli classes. For instance,

$$F = \{1[x \in S] : S \subset \mathbb{R}, |S| < \infty\}.$$

Then for a continuous distribution P, $Pf = 0$ for any $f \in F$, but $\sup_{f \in F} P_n f = 1$ for all n. So although $P_n f \xrightarrow{as} Pf$ for all $f \in F$, this convergence is not uniform over F. F is too large.

Outline

- Uniform laws of large numbers
- Rademacher complexity and uniform laws
  (Concentration. Symmetrization. Restrictions.)
- Controlling Rademacher complexity:
  - Growth function
  - VC-dimension
  - Structural results for Rademacher complexity
- Neural networks
  - VC-dimension
  - Large margin classifiers
  - Rademacher averages for sigmoid networks
  - Rademacher averages for ReLU networks
- Interpolating prediction rules

Uniform Laws and Rademacher Complexity

We'll look at a proof of a uniform law of large numbers that involves two steps:

1. Concentration of $\|P - P_n\|_F$ about its expectation.
2. Symmetrization, which bounds $\mathbb{E}\|P - P_n\|_F$ in terms of the Rademacher complexity of F, $\mathbb{E}\|R_n\|_F$.

Uniform Laws and Rademacher Complexity

Definition

The **Rademacher complexity** of F is $\mathbb{E}\|R_n\|_F$, where the empirical process R_n is defined as

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(X_i),$$

and the $\epsilon_1, \ldots, \epsilon_n$ are Rademacher random variables: i.i.d. uniform on $\{\pm 1\}$.

Note that this is the expected supremum of the alignment between the random $\{\pm 1\}$-vector ϵ and $F(X_1^n)$, the set of n-vectors obtained by restricting F to the sample $X_1, \ldots, X_n$.

Uniform Laws and Rademacher Complexity

Theorem

For any F, $\mathbb{E}\|P - P_n\|_F \leq 2\mathbb{E}\|R_n\|_F$.
If $F \subset [0,1]^{\mathcal{X}}$,

$$\frac{1}{2}\mathbb{E}\|R_n\|_F - \sqrt{\frac{\log 2}{2n}} \leq \mathbb{E}\|P - P_n\|_F \leq 2\mathbb{E}\|R_n\|_F,$$

and, with probability at least $1 - 2\exp(-2\epsilon^2 n)$,

$$\mathbb{E}\|P - P_n\|_F - \epsilon \leq \|P - P_n\|_F \leq \mathbb{E}\|P - P_n\|_F + \epsilon.$$

Thus, $\mathbb{E}\|R_n\|_F \to 0$ iff $\|P - P_n\|_F \stackrel{as}{\to} 0$.

That is, the supremum of the empirical process $P - P_n$ is concentrated about its expectation, and its expectation is about the same as the expected sup of the Rademacher process R_n.

Uniform Laws and Rademacher Complexity

The first step is to symmetrize by replacing Pf by $P'_n f = \frac{1}{n}\sum_{i=1}^n f(X'_i)$. In particular, we have

$$\begin{aligned}
\mathbb{E}\|P - P_n\|_F &= \mathbb{E}\sup_{f\in F}\left|\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (f(X'_i) - f(X_i))\middle| X_1^n\right]\right| \\
&\leq \mathbb{E}\mathbb{E}\left[\sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^n (f(X'_i) - f(X_i))\right|\middle| X_1^n\right] \\
&= \mathbb{E}\sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^n (f(X'_i) - f(X_i))\right| = \mathbb{E}\|P'_n - P_n\|_F.
\end{aligned}$$

Uniform Laws and Rademacher Complexity

Another symmetrization: for any $\epsilon_i \in \{\pm 1\}$,

$$\begin{aligned}\mathbb{E}\|P_n' - P_n\|_F &= \mathbb{E}\sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^n (f(X_i') - f(X_i))\right| \\ &= \mathbb{E}\sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i (f(X_i') - f(X_i))\right|,\end{aligned}$$

This follows from the fact that X_i and X_i' are i.i.d., and so the distribution of the supremum is unchanged when we swap them. And so in particular the expectation of the supremum is unchanged. And since this is true for any ϵ_i, we can take the expectation over any random choice of the ϵ_i. We'll pick them independently and uniformly.

Uniform Laws and Rademacher Complexity

$$
\begin{aligned}
&\mathbb{E} \sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i (f(X_i') - f(X_i)) \right| \\
&\leq \mathbb{E} \sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i') \right| + \sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right| \\
&= 2 \underbrace{\mathbb{E} \sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right|}_{\text{Rademacher complexity}} = 2\mathbb{E}\|R_n\|_F,
\end{aligned}
$$

where R_n is the *Rademacher process* $R_n(f) = (1/n) \sum_{i=1}^{n} \epsilon_i f(X_i)$.

Uniform Laws and Rademacher Complexity

The second inequality (*desymmetrization*) follows from:

$$
\begin{aligned}
\mathbb{E}\|R_n\|_F &\leq \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\left(f(X_i)-\mathbb{E}f(X_i)\right)\right\|_F + \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\mathbb{E}f(X_i)\right\|_F \\
&\leq \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\left(f(X_i)-f(X_i')\right)\right\|_F + \|P\|_F\,\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right| \\
&= \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(f(X_i)-\mathbb{E}f(X_i)+\mathbb{E}f(X_i')-f(X_i')\right)\right\|_F \\
&\qquad + \|P\|_F\,\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right| \\
&\leq 2\mathbb{E}\,\|P_n-P\|_F + \sqrt{\frac{2\log 2}{n}}.
\end{aligned}
$$

Uniform Laws and Rademacher Complexity

Next, since $f(X_i) \in [0, 1]$, we have that the following function of the random variables $X_1, \ldots, X_n$ satisfies a bounded differences property with bound $1/n$:

$$\sup_{f \in F} |Pf - P_n f|.$$

The *bounded differences inequality* implies that, with probability at least $1 - \exp(-2\epsilon^2 n)$,

$$\|P - P_n\|_F \leq \mathbb{E}\|P - P_n\|_F + \epsilon.$$

Bounded Differences Inequality

Theorem [Bounded differences inequality]

Suppose $f : \mathcal{X}^n \to \mathbb{R}$ satisfies the following **bounded differences property**: for all $x_1, \ldots, x_n, x_i' \in \mathcal{X}$,

$$\left| f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n) \right| \leq B_i.$$

Then

$$P\left(|f(X) - \mathbb{E}f(X)| \geq t \right) \leq 2 \exp\left(-\frac{2t^2}{\sum_i B_i^2} \right).$$

Uniform Laws and Rademacher Complexity

Theorem

For any F, $\mathbb{E}\|P - P_n\|_F \leq 2\mathbb{E}\|R_n\|_F$.
If $F \subset [0,1]^{\mathcal{X}}$,

$$\frac{1}{2}\mathbb{E}\|R_n\|_F - \sqrt{\frac{\log 2}{2n}} \leq \mathbb{E}\|P - P_n\|_F \leq 2\mathbb{E}\|R_n\|_F,$$

and, with probability at least $1 - 2\exp(-2\epsilon^2 n)$,

$$\mathbb{E}\|P - P_n\|_F - \epsilon \leq \|P - P_n\|_F \leq \mathbb{E}\|P - P_n\|_F + \epsilon.$$

Thus, $\mathbb{E}\|R_n\|_F \to 0$ iff $\|P - P_n\|_F \stackrel{as}{\to} 0$.

That is, the supremum of the empirical process $P - P_n$ is concentrated about its expectation, and its expectation is about the same as the expected sup of the Rademacher process R_n.

Outline

- Uniform laws of large numbers
- Rademacher complexity and uniform laws
  (Concentration. Symmetrization. Restrictions.)
- Controlling Rademacher complexity:
  - Growth function
  - VC-dimension
  - Structural results for Rademacher complexity
- Neural networks
  - VC-dimension
  - Large margin classifiers
  - Rademacher averages for sigmoid networks
  - Rademacher averages for ReLU networks
- Interpolating prediction rules

Controlling Rademacher Complexity

So how do we control $\mathbb{E}\|R_n\|_F$? There are several approaches:

1. $|F(X_1^n)|$ small. ($\max|F(x_1^n)|$ is the **growth function**)
2. For binary-valued functions: Vapnik-Chervonenkis dimension. Bounds rate of growth function. Can be bounded for parameterized families.
3. Structural results on Rademacher complexity: Obtaining bounds for function classes constructed from other function classes.
4. Covering numbers. Dudley entropy integral, Sudakov lower bound.
5. For real-valued functions: scale-sensitive dimensions.

Controlling Rademacher Complexity: Growth Function

For the class of distribution functions, $G = \{x \mapsto 1[x \leq \alpha] : \alpha \in \mathbb{R}\}$, the set of restrictions,

$$G(x_1^n) = \{(g(x_1), \ldots, g(x_n)) : g \in G\}$$

is always small: $|G(x_1^n)| \leq \Pi_G(n) = n + 1$.

Definition

For a class $F \subseteq \{0,1\}^{\mathcal{X}}$, the **growth function** is

$$\Pi_F(n) = \max\{|F(x_1^n)| : x_1, \ldots, x_n \in \mathcal{X}\}.$$

Controlling Rademacher Complexity: Growth Function

Lemma

For $f \in F$ satisfying $|f(x)| \leq 1$,

$$\begin{aligned}\mathbb{E}\|R_n\|_F &\leq \mathbb{E}\sqrt{\frac{2\log(|F(X_1^n) \cup -F(X_1^n)|)}{n}} \\ &\leq \sqrt{\frac{2\log(2\mathbb{E}|F(X_1^n)|)}{n}} \\ &\leq \sqrt{\frac{2\log(2\Pi_F(n)}{n}},\end{aligned}$$

where R_n is the Rademacher process:

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(X_i).$$

and $F(X_1^n)$ is the set of restrictions of functions in F to $X_1, \ldots, X_n$.

Controlling Rademacher Complexity: Growth Function

Proof: For $A \subseteq \mathbb{R}^n$ with $R^2 = \frac{\max_{a \in A} \|a\|_2^2}{n}$, we have that

$$\mathbb{E} \sup_{a \in A} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i a_i \right| \leq \sqrt{\frac{2R^2 \log(|A \cup -A|)}{n}}.$$

Here, we have $A = F(X_1^n)$, so $R \leq 1$, and we get

$$\begin{aligned}
\mathbb{E}\|R_n\|_F &= \mathbb{E}\mathbb{E}\left[\|R_n\|_{F(X_1^n)} | X_1, \ldots, X_n\right] \\
&\leq \mathbb{E}\sqrt{\frac{2\log(2|F(X_1^n)|)}{n}} \\
&\leq \sqrt{\frac{2\mathbb{E}\log(2|F(X_1^n)|)}{n}} \\
&\leq \sqrt{\frac{2\log(2\mathbb{E}|F(X_1^n)|)}{n}}.
\end{aligned}$$

Finite Class Lemma

We used the following result.

Lemma [Finite Classes]

For $A \subseteq \mathbb{R}^n$ with $R^2 = \dfrac{\max_{a \in A} \|a\|_2^2}{n}$,

$$\mathbb{E} \sup_{a \in A} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i a_i \leq \sqrt{\frac{2R^2 \log |A|}{n}}.$$

Hence

$$\mathbb{E} \sup_{a \in A} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i a_i \right| = \mathbb{E} \sup_{a \in A \cup -A} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i a_i \leq \sqrt{\frac{2R^2 \log(2|A|)}{n}}.$$

Finite Class Lemma

Proof:

$$\exp\left(\lambda \mathbb{E} \sup_a \frac{1}{n}\sum_{i=1}^n \epsilon_i a_i\right) \leq \mathbb{E}\exp\left(\lambda \sup_a \frac{1}{n}\sum_{i=1}^n \epsilon_i a_i\right)$$
$$= \mathbb{E}\sup_a \exp\left(\lambda \frac{1}{n}\sum_{i=1}^n \epsilon_i a_i\right)$$
$$\leq \sum_a \mathbb{E}\exp\left(\lambda \frac{1}{n}\sum_{i=1}^n \epsilon_i a_i\right)$$
$$\leq \sum_a \exp\left(\frac{\lambda^2 \|a\|_2^2}{2n^2}\right)$$
$$\leq |A| \exp\left(\frac{\lambda^2 R^2}{2n}\right),$$

using the fact that $\epsilon_i a_i / n$ is bounded, hence sub-Gaussian. Picking $\lambda^2 = 2n \log |A| / R^2$ gives the result.

Concentration of Sub-Gaussian Random Variables

Definition

X is **sub-Gaussian** with parameter σ^2 if, for all $\lambda \in \mathbb{R}$,

$$\log M_{X-\mu}(\lambda) \leq \frac{\lambda^2 \sigma^2}{2},$$

where $M_{X-\mu}(\lambda) = \mathbb{E} \exp(\lambda(X - \mu))$ (for $\mu = \mathbb{E}X$) is the **moment-generating function** of $X - \mu$.

- Examples: X Gaussian; X a.s. bounded.
- A sum of independent sub-Gaussian random variables is sub-Gaussian; the parameters add.
- Chernoff bound for X sub-Gaussian implies

$$P(|X - \mu| \geq t) \leq 2 \exp(-t^2/(2\sigma^2)).$$

Controlling Rademacher Complexity: Growth Function

e.g. For the class of distribution functions, $G = \{x \mapsto 1[x \geq \alpha] : \alpha \in \mathbb{R}\}$, we saw that $|G(x_1^n)| \leq n + 1$. So $\mathbb{E}\|R_n\|_F \leq \sqrt{\frac{2 \log 2(n+1)}{n}}$.
e.g. F parameterized by k bits:
If $F = \left\{x \mapsto g(x, \theta) : \theta \in \{0,1\}^k\right\}$ for some $g : \mathcal{X} \times \{0,1\}^k \to [0,1]$,

$$|F(x_1^n)| \leq 2^k,$$

$$\mathbb{E}\|R_n\|_F \leq \sqrt{\frac{2(k+1)\log 2}{n}}.$$

Notice that $\mathbb{E}\|R_n\|_F \to 0$.

Growth Function

Definition

For a class $F \subseteq \{0,1\}^{\mathcal{X}}$, the **growth function** is

$$\Pi_F(n) = \max\{|F(x_1^n)| : x_1, \ldots, x_n \in \mathcal{X}\}.$$

- $\mathbb{E}\|R_n\|_F \leq \sqrt{\frac{2\log(2\Pi_F(n))}{n}}$.
- $\Pi_F(n) \leq |F|$, $\lim_{n\to\infty} \Pi_F(n) = |F|$.
- $\Pi_F(n) \leq 2^n$. (But then this gives no useful bound on $\mathbb{E}\|R_n\|_F$.)
- Notice that $\log \Pi_F(n) = o(n)$ implies $\mathbb{E}\|R_n\|_F \to 0$.

Outline

- Uniform laws of large numbers
- Rademacher complexity and uniform laws
  (Concentration. Symmetrization. Restrictions.)
- Controlling Rademacher complexity:
  - Growth function
  - VC-dimension
  - Structural results for Rademacher complexity
- Neural networks
  - VC-dimension
  - Large margin classifiers
  - Rademacher averages for sigmoid networks
  - Rademacher averages for ReLU networks
- Interpolating prediction rules

Vapnik-Chervonenkis Dimension

Definition

A class $F \subseteq \{0,1\}^{\mathcal{X}}$ **shatters** $\{x_1, \ldots, x_d\} \subseteq \mathcal{X}$ means that $|F(x_1^d)| = 2^d$.
The Vapnik-Chervonenkis dimension of F is

$$\begin{aligned} d_{VC}(F) &= \max\{d : \text{some } x_1, \ldots, x_d \in \mathcal{X} \text{ is shattered by } F\} \\ &= \max\left\{d : \Pi_F(d) = 2^d\right\}. \end{aligned}$$

Vapnik-Chervonenkis Dimension: "Sauer's Lemma"

Theorem [Vapnik-Chervonenkis]

$d_{VC}(F) \leq d$ implies

$$\Pi_F(n) \leq \sum_{i=0}^{d} \binom{n}{i}.$$

If $n \geq d$, the latter sum is no more than $\left(\frac{en}{d}\right)^d$.

So the VC-dimension is a single integer summary of the growth function: either it is finite, and $\Pi_F(n) = O(n^d)$, or $\Pi_F(n) = 2^n$. No other growth is possible.

$$\Pi_F(n) \begin{cases} = 2^n & \text{if } n \leq d, \\ \leq (e/d)^d \, n^d & \text{if } n > d. \end{cases}$$

Thus, for $d_{VC}(F) \leq d$ and $n \geq d$, we have

$$\mathbb{E}\|R_n\|_F \leq \sqrt{\frac{2\log(2\Pi_F(n))}{n}} \leq \sqrt{\frac{2\log 2 + 2d\log(en/d)}{n}}.$$

VC-dimension Bounds for Parameterized Families

Consider a parameterized class of binary-valued functions,

$$F = \{x \mapsto f(x, \theta) : \theta \in \mathbb{R}^p\},$$

where $f : \mathbb{R}^m \times \mathbb{R}^p \to \{\pm 1\}$.
Suppose that f can be computed using no more than t operations of the following kinds:

1. arithmetic ($+$, $-$, $\times$, $/$),
2. comparisons ($>$, $=$, $<$),
3. output ± 1.

Theorem [Goldberg and Jerrum]

$d_{VC}(F) \leq 4p(t+2)$.

Outline

- Uniform laws of large numbers
- Rademacher complexity and uniform laws
  (Concentration. Symmetrization. Restrictions.)
- Controlling Rademacher complexity:
  - Growth function
  - VC-dimension
  - Structural results for Rademacher complexity
- Neural networks
  - VC-dimension
  - Large margin classifiers
  - Rademacher averages for sigmoid networks
  - Rademacher averages for ReLU networks
- Interpolating prediction rules

Rademacher Complexity: Structural Results

Theorem

1. $F \subseteq G$ implies $\|R_n\|_F \leq \|R_n\|_G$.
2. $\|R_n\|_{cF} = |c|\|R_n\|_F$.
3. For $|g(X)| \leq 1$, $|\mathbb{E}\|R_n\|_{F+g} - \mathbb{E}\|R_n\|_F| \leq \sqrt{2\log 2/n}$.
4. $\|R_n\|_{\mathrm{co}F} = \|R_n\|_F$, where $\mathrm{co}F$ is the convex hull of F.
5. If $\phi : \mathbb{R} \times \mathcal{Z}$ has $\alpha \mapsto \phi(\alpha, z)$ 1-Lipschitz for all z and $\phi(0, z) = 0$, then for $\phi(F) = \{z \mapsto \phi(f(z), z)\}$, $\mathbb{E}\|R_n\|_{\phi(F)} \leq 2\mathbb{E}\|R_n\|_F$.

Rademacher Complexity Structural Results: Proofs

(1) and (2) are immediate. For (3):

$$\|R_n\|_{F+g} = \sup_{f \in F} \left| \frac{1}{n} \sum_i \epsilon_i \left(f(X_i) + g(X_i) \right) \right|,$$

$$\text{so} \qquad |\mathbb{E}\|R_n\|_{F+g} - \mathbb{E}\|R_n\|_F| \leq \mathbb{E}\,|R_n(g)| \leq \sqrt{\frac{2 \log 2}{n}}$$

for $|g(X)| \leq 1$.
(4) follows from the fact that a linear criterion in a convex set is maximized at an extreme point.
(5) is due to Ledoux and Talagrand, and has an elementary proof.

Uniform Laws of Large Numbers: Summary

Rademacher complexity

Rademacher complexity characterizes uniform laws: For $F \subset [0,1]^{\mathcal{X}}$,

$$\frac{1}{2}\mathbb{E}\|R_n\|_F - \sqrt{\frac{\log 2}{2n}} \leq \mathbb{E}\|P - P_n\|_F \leq 2\mathbb{E}\|R_n\|_F,$$

and, with probability at least $1 - 2\exp(-2\epsilon^2 n)$,

$$\mathbb{E}\|P - P_n\|_F - \epsilon \leq \|P - P_n\|_F \leq \mathbb{E}\|P - P_n\|_F + \epsilon.$$

Vapnik-Chervonenkis dimension

Uniform convergence uniformly over probability distributions is equivalent to finiteness of the VC-dimension.

Outline

- Uniform laws of large numbers
- Rademacher complexity and uniform laws
  (Concentration. Symmetrization. Restrictions.)
- Controlling Rademacher complexity:
  - Growth function
  - VC-dimension
  - Structural results for Rademacher complexity
- Neural networks
  - VC-dimension
  - Large margin classifiers
  - Rademacher averages for sigmoid networks
  - Rademacher averages for ReLU networks
- Interpolating prediction rules