Supervised Learning (COMP0078)

7. Learning Theory (Part II): Rademacher Complexity

Carlo Ciliberto

University College London
Department of Computer Science

Outline

- Rademacher Complexity and Generalization Error
- Contraction Lemma
- Rademacher complexity in practice
- Constrained optimization

Last Class

We introduced regularization as the abstract strategy of controlling the expressiveness of an estimator, via a parameter $\gamma \geq 0$.

$$f_{n,\gamma} = \underset{f \in \mathcal{H}_\gamma}{\arg\min}\ \mathcal{E}_n(f) \qquad\qquad f_\gamma = \underset{f \in \mathcal{H}_\gamma}{\arg\min}\ \mathcal{E}(f)$$

Assuming $\mathcal{H} = \bigcup_{\gamma \geq 0} \mathcal{H}_\gamma$ we derived the bias-variance decomposition of the excess risk

$$\begin{aligned} &\mathcal{E}(f_{n,\gamma}) - \mathcal{E}(f_*) \\ &\qquad = \underbrace{\mathcal{E}(f_{n,\gamma}) - \mathcal{E}(f_\gamma)}_{\text{sample error/variance}} + \underbrace{\mathcal{E}(f_\gamma) - \inf_{f \in \mathcal{H}} \mathcal{E}(f)}_{\text{approximation/bias}} + \underbrace{\inf_{f \in \mathcal{H}} \mathcal{E}(f) - \mathcal{E}(f_*)}_{\text{irreducible error}} \end{aligned}$$

Depending on the number n of training points available, the ERM strategy would choose $\gamma = \gamma(n)$ as striking the best balance between variance and bias errors.

Last Class - Excess Risk Decomposition

- The irreducible error depends on our choice of $\mathcal{H}$ with respect to the learning problem. We can not say much about it except if we choose $\mathcal{H}$ to be "universal" (several choices available), which guarantees the irreducible error to be 0!

- The approximation error depends on the learning problem, the space $\mathcal{H}$ and our choice of regularization γ.
  - We will get back to this once we will start discussing some more concrete implementations of the regularization strategy (actual algorithms).
  - Since we don't know ρ, we will always need to make some assumptions to say something meaningful about it!

- The sample error depends on how much "freedom" the space $\mathcal{H}_\gamma$ has, as a function of γ.

Last Class - Generalization Error

We have seen how generalization error is a key quantity to control if we want to study the sample error of our learning algorithm...

$$\mathcal{E}(f_{n,\gamma}) - \mathcal{E}_n(f_{n,\gamma})$$

This follows by decomposing the sample error

$$\begin{aligned}
&\mathcal{E}(f_{n,\gamma}) - \mathcal{E}(f_\gamma) \\
&\quad = \underbrace{\mathcal{E}(f_{n,\gamma}) - \mathcal{E}_n(f_{n,\gamma})}_{\text{Generalization error}} + \underbrace{\mathcal{E}_n(f_{n,\gamma}) - \mathcal{E}_n(f_\gamma)}_{\leq 0} + \underbrace{\mathcal{E}_n(f_\gamma) - \mathcal{E}(f_\gamma)}_{\substack{0 \text{ in expectation} \\ O(1/\sqrt{n}) \text{ in probability}}}
\end{aligned}$$

Last Class - Finite Hypotheses Spaces

We observed that limiting ourselves to hypotheses spaces containing a finite number of functions, we could control the generalization error,

$$\mathcal{E}(f_{n,\gamma}) - \mathcal{E}_n(f_{n,\gamma}) \leq |\mathcal{H}_\gamma| \sqrt{\frac{V_\gamma}{n}} \qquad \text{with} \qquad V_\gamma = \sup_{f \in \mathcal{H}_\gamma} \operatorname{Var} \ell(f(x), y).$$

- **Pros.** Plugging this result in the excess risk decomposition we are able to actually study the predicton performance of the learning algorithm $f_{n,\gamma}$.
- **Cons.** The cardinality $|\mathcal{H}_\gamma|$ is *very* concerning: even if we have seen that from a statistical perspective we can mitigate its effect (e.g. using Hoeffding's inequality and make it appear as a logaritmic factor), solving ERM on $\mathcal{H}_\gamma$ requires evaluating the empirical risk $|\mathcal{H}_\gamma|$ times!

Beyond Finite Hypotheses Spaces

Ideally... We would like to find suitable spaces $\mathcal{H}_\gamma$ such that:

1. Any algorithm (ERM included) producing functions in $\mathcal{H}_\gamma$ enjoys good generalization bounds (like it is the case for finite spaces).
2. Solving ERM (or in any case, carrying out the required optimization) over $\mathcal{H}_\gamma$ is efficient with respect to *n* and γ (e.g. can be done in polynomial time).
3. The family of $\{\mathcal{H}_\gamma\}_{\gamma>0}$ is “fast” in approximating $\mathcal{H}$. More precisely, under weak assumptions on the learning problem, the bias-variance trade-off identified by γ should yield to fast learning rates.

Beyond Finite Hypotheses Spaces (Continued)

Let's look back at the way we were able to control the generalization of f_n over a finite space of Hypotheses $\mathcal{H}$.

$$\begin{aligned}
\mathbb{E}[\mathcal{E}(f_n) - \mathcal{E}_n(f_n)] &\leq \mathbb{E}\left[\sup_{f \in \mathcal{H}} \mathcal{E}(f) - \mathcal{E}_n(f)\right] \\
&\leq \sum_{f \in \mathcal{H}} \mathbb{E}\, \mathcal{E}(f) - \mathcal{E}_n(f) \\
&\leq |\mathcal{H}| \sqrt{\frac{V_{\mathcal{H}}}{n}}.
\end{aligned}$$

Both inequalities first and second inequalities are possibly loose, but the second one, replacing the $\sup$ with the sum over all possible functions in $\mathcal{H}$ is arguably the worst...

...can we do better to control $\mathbb{E}\left[\sup_{f \in \mathcal{H}} \mathcal{E}(f) - \mathcal{E}_n(f)\right]$?

Yes, for example using Rademacher complexity.

Rademacher Complexity

Rademacher complexity is a way to measure how expressive a family of hypotheses is, by measuring how "well" the functions it contains correlate with random noise.

Empirical Rademacher Complexity: Let $\mathcal{Z}$ be a set and $S = (z_i)_{i=1}^n$ a dataset on $\mathcal{Z}$. The *empirical Rademacher complexity* of a space of hypotheses $\mathcal{H}\{f : \mathcal{Z} \to \mathbb{R}\}$ is

$$\mathcal{R}_S(\mathcal{H}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^n \sigma_i f(z_i) \right]$$

where $\sigma = (\sigma_i)_{i=1}^n$ and σ_i uniformly sampled in $\{-1, 1\}$ (independent to each other) known as *Rademacher variables*.

Rademacher Complexity: Let ρ be a probability measure on $\mathcal{Z}$

$$\mathcal{R}_n(\mathcal{H}) = \mathcal{R}_{\rho,n}(\mathcal{H}) = \mathbb{E}_{S \sim \rho^n}[\mathcal{R}_S(\mathcal{H})]$$

Rademacher Complexity

Rademacher complexity is a well-established measure that:

- Can be controlled for a large number of popular spaces of hypotheses (we will see one key example in a bit).
- Is related to many other complexity measures:[1]
    - Covering numbers,
    - Gaussian complexity,
    - Growth function,
    - Vapnik-Chervonenkis (VC) dimension,
    - …

...and it is *very* reminiscent of term we would like to control (an expectation of a sup)...

...could we use it to upper bound $\mathbb{E}\left[\sup_{f \in \mathcal{H}} \mathcal{E}(f) - \mathcal{E}_n(f)\right]$?

[1] e.g. can upper bound and/or be upper bounded by

Rademacher Complexity and Generalization Error

We will try now to connect the “worst” generalization error over $\mathcal{H}$ and the Rademacher complexity of $\mathcal{H}$.

In particular, we will show that

$$\mathbb{E}\left[\sup_{f\in\mathcal{H}} \mathcal{E}(f) - \mathcal{E}_n(f)\right] \leq 2\mathcal{R}_n(\ell \circ \mathcal{H})$$

where

$$\ell \circ \mathcal{H} = \{g(x,y) = \ell(f(x),y) \mid f \in \mathcal{H}\}$$

Let’s prove this...

Back to the "worst" generalization error

Notation. For clarity, in the following we denote the empirical risk of a function f with respect to a dataset $S \sim \rho^n$ as $\mathcal{E}_S(f)$.

Recall that for any dataset $S \sim \rho^n$, the expectation of the empirical risk corresponds to the expected risk, namely $\mathbb{E}_S\, \mathcal{E}_S(f) = \mathcal{E}(f)$. Then, by introducing a new "virtual" dataset $S' \sim \rho^n$, we have

$$\mathbb{E}_S \left[\sup_{f \in \mathcal{H}} \mathcal{E}(f) - \mathcal{E}_S(f)\right] = \mathbb{E}_S \left[\sup_{f \in \mathcal{H}} \mathbb{E}_{S'} \left[\mathcal{E}_{S'}(f) - \mathcal{E}_S(f)\right]\right]$$

Moreover, since the the sup function is convex, we have

$$\mathbb{E}_S \left[\sup_{f \in \mathcal{H}} \mathbb{E}_{S'} \left[\mathcal{E}_{S'}(f) - \mathcal{E}_S(f)\right]\right] \leq \mathbb{E}_{S,S'} \left[\sup_{f \in \mathcal{H}} \mathcal{E}_{S'}(f) - \mathcal{E}_S(f)\right]$$

Introducing the Rademacher Variables

Let $S = (x_i, y_i)_{i=1}^n$ and $S' = (x_i', y_i')_{i=1}^n$, then

$$\mathcal{E}_{S'}(f) - \mathcal{E}_S(f) = \frac{1}{n}\sum_{i=1}^n \big(\ell(f(x_i'), y_i') - \ell(f(x_i), y_i)\big).$$

Introduce the Rademacher variables σ_i sampled with uniform probability in $\{-1, 1\}$. We note the following equality holds

$$\begin{aligned}\mathbb{E}_{S,S'}&\left[\sup_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \big(\ell(f(x_i'), y_i') - \ell(f(x_i), y_i)\big)\right] \\ &= \mathbb{E}_{S,S',\sigma}\left[\sup_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \sigma_i \big(\ell(f(x_i'), y_i') - \ell(f(x_i), y_i)\big)\right].\end{aligned}$$

Why?

Introducing the Rademacher Variables

Let $S = (x_i, y_i)_{i=1}^n$ and $S' = (x_i', y_i')_{i=1}^n$, then

$$\mathcal{E}_{S'}(f) - \mathcal{E}_S(f) = \frac{1}{n}\sum_{i=1}^n \big(\ell(f(x_i'), y_i') - \ell(f(x_i), y_i)\big).$$

Introduce the Rademacher variables σ_i sampled with uniform probability in $\{-1, 1\}$. We note the following equality holds

$$\mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \big(\ell(f(x_i'), y_i') - \ell(f(x_i), y_i)\big)\right]$$
$$= \mathbb{E}_{S,S',\sigma}\left[\sup_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \sigma_i \big(\ell(f(x_i'), y_i') - \ell(f(x_i), y_i)\big)\right].$$

Why? Because sampling $\sigma_i = -1$ can be interpreted as "swapping" the sample (x_i, y_i) from S with the (x_i', y_i') in S'. But the expectation is considering all possible combinations of S and S'. We are only changing the order of elements in the expectation but not the result.

Sub-additivity of the Supremum

By recalling that the supremum is *sub-additve*, namely $\sup_x f(x) + g(x) \leq \sup_x f(x) + \sup_x g(x)$, we have

$$\begin{aligned}
\mathbb{E}_{S,S',\sigma} \sup_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i\big(\ell(f(x_i'), y_i') - \ell(f(x_i), y_i)\big) \\
&\leq \mathbb{E}_{S',\sigma} \sup_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(f(x_i'), y_i') + \mathbb{E}_{S,\sigma} \sup_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} -\sigma_i \ell(f(x_i), y_i) \\
&\leq 2\ \mathbb{E}_{S,\sigma} \sup_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(f(x_i), y_i).
\end{aligned}$$

The last inequality follows by observing that the σ_is absorb any change of sign and S and S' play the same role in the two elements in the sum.

Back to the Rademacher Complexity

The last term we got is actually a Rademacher complexity...

To see it, consider

$$\mathcal{G} = \{g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \mid g(x, y) = \ell(f(x), y) \ \exists f \in \mathcal{H}\}$$

We denote $\mathcal{G} = \ell \circ \mathcal{H}$ as the set of functions obtained by composing the loss ℓ with the hypotheses in $\mathcal{H}$. Then,

$$\mathbb{E}_{S,\sigma} \sup_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(f(x_i), y_i) = \mathbb{E}_{S,\sigma} \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(\underbrace{z_i}_{(x_i, y_i)}) = \mathcal{R}_n(\ell \circ \mathcal{H})$$

Bringing everything back together, we have

$$\mathbb{E} \left[\sup_{f \in \mathcal{H}} \mathcal{E}(f) - \mathcal{E}_n(f) \right] \leq 2\mathcal{R}_n(\ell \circ \mathcal{H})$$

as required.

Dependency on the Loss

We were able to bound the generalization bound in terms of $\mathcal{R}_n(\ell \circ \mathcal{H})$.

However, in practice, we can expect to have results characterizing the Rademacher complexity of a space $\mathcal{H}$ for some well-established hypotheses space (and we will see some of them below)...

Question. Can we control $\mathcal{R}_n(\ell \circ \mathcal{H})$ in terms of $\mathcal{R}_n(\mathcal{H})$?

Yes! Provided we make some assumptions on the loss...

Contraction Lemma

We have seen that most we use have appealing properties, e.g. smoothness, convexity, Lipschitz, etc...

Lemma (Contraction). *Let $\ell(\cdot, y)$ be L-Lipschitz uniformly for $y \in \mathcal{Y}$ with $L > 0$. Then, for any set $S = (x_i, y_i)_{i=1}^n$*

$$\mathcal{R}_S(\ell \circ \mathcal{H}) \leq L\mathcal{R}_{S_\mathcal{X}}(\mathcal{H}),$$

with $S_\mathcal{X} = (x_i)_{i=1}^n$. Furthermore, for any ρ probability distribution on $\mathcal{X} \times \mathcal{Y}$ and any $n \in \mathbb{N}$,

$$\mathcal{R}_n(\ell \circ \mathcal{H}) \leq L\mathcal{R}_n(\mathcal{H}).$$

Contraction Lemma

Let us start by isolating the contribution of the term $\sigma_1 \ell(f(x_1), y_1)$ in the Rademacher complexity:

$$
\begin{aligned}
\mathcal{R}_S(\ell \circ \mathcal{H}) &= \mathbb{E}_\sigma \sup_{f \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i \ell(f(x_i), y_i) \\
&= \mathbb{E}_\sigma \sup_{f \in \mathcal{H}} \left[\sigma_1 \ell(f(x_1), y_1) + \sum_{i=2}^{n} \sigma_i \ell(f(x_i), y_i) \right] \\
&= \frac{1}{2} \mathbb{E}_{\sigma_2, \ldots \sigma_n} \sup_{f \in \mathcal{H}} \left[\ell(f(x_1), y_1) + \sum_{i=2}^{n} \sigma_i \ell(f(x_i), y_i) \right] \\
&\quad + \frac{1}{2} \mathbb{E}_{\sigma_2, \ldots \sigma_n} \sup_{f \in \mathcal{H}} \left[-\ell(f(x_1), y_1) + \sum_{i=2}^{n} \sigma_i \ell(f(x_i), y_i) \right]
\end{aligned}
$$

where we have explicitly written the expectation with respect to σ_1 (which is uniformly sampled from $\{-1, 1\}$).

Contraction Lemma (Cont.)

By considering the supremum over two functions f and f', we then have

$$\mathcal{R}_S(\ell \circ \mathcal{H}) = \frac{1}{2}\mathbb{E}_{\sigma_2,\dots,\sigma_n} \sup_{f,f' \in \mathcal{H}} \Bigg[\ell(f(x_1), y_1) - \ell(f'(x_1), y_1) + \sum_{i=2}^{n} \sigma_i \ell(f(x_i), y_i) + \sum_{i=2}^{n} \sigma_i \ell(f'(x_i), y_i)\Bigg].$$

Since the loss is L-Lipschiz...

$$\mathcal{R}_S(\ell \circ \mathcal{H}) \leq \frac{1}{2}\mathbb{E}_{\sigma_2,\dots,\sigma_n} \sup_{f,f' \in \mathcal{H}} \Bigg[L|f(x_1) - f'(x_1)| + \sum_{i=2}^{n} \sigma_i \ell(f(x_i), y_i) + \sum_{i=2}^{n} \sigma_i \ell(f'(x_i), y_i)\Bigg]$$

Contraction Lemma (Cont.)

Since f and f' are from the same set $\mathcal{H}$ and the last two terms are identical functions of f or f', we can remove the absolute value, namely

$$\frac{1}{2} \sup_{f,f' \in \mathcal{H}} \left[L|f(x_1) - f'(x_1)| + \sum_{i=2}^{n} \sigma_i \ell(f(x_i), y_i) + \sum_{i=2}^{n} \sigma_i \ell(f'(x_i), y_i) \right]$$

$$= \frac{1}{2} \sup_{f,f' \in \mathcal{H}} \left[L(f(x_1) - f'(x_1)) + \sum_{i=2}^{n} \sigma_i \ell(f(x_i), y_i) + \sum_{i=2}^{n} \sigma_i \ell(f'(x_i), y_i) \right]$$

By splitting again the supremum with respect to f and f', we can write everything as

$$= \mathbb{E}_{\sigma_1} \sup_{f \in \mathcal{H}} \left[L\sigma_1 f(x_1) + \sum_{i=2}^{n} \sigma_i \ell(f(x_i), y_i) \right]$$

Contraction Lemma (Cont.)

Repeating the same argument for $i = 2, \ldots, n$, we conclude that

$$\begin{aligned}\mathcal{R}_S(\ell \circ \mathcal{H}) &= \mathbb{E}_\sigma \sup_{f \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i \ell(f(x_i), y_i) \\ &\leq L \mathbb{E}_\sigma \sup_{f \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i f(x_i) = L \mathcal{R}_{S_\mathcal{X}}(\mathcal{H}),\end{aligned}$$

as desired. The result for the (expected) Rademacher complexity

$$\mathcal{R}_n(\ell \circ \mathcal{H}) \leq L\, \mathcal{R}_n(\mathcal{H}),$$

follows by taking the expectation with respect to $S \sim \rho^n$.

Bringing everything together

Therefore, by assuming ℓ to be L-lipschitz, we can control the worst generalization error as

$$\mathbb{E}\mathcal{E}(f_n) - \mathcal{E}_n(f_n) \leq \mathbb{E} \sup_{f \in \mathcal{H}} \mathcal{E}(f) - \mathcal{E}_n(f) \leq 2L\,\mathcal{R}(\mathcal{H})$$

Can we control the same result in probability?

McDiarmid Inequality

Theorem. *Let $\mathcal{Z}$ be a set and $g : \mathcal{Z}^n \to \mathbb{R}$ be a function such that there exists $c > 0$ such that for any $i = 1, \ldots, n$ and any $z_1, \ldots, z_n, z_i' \in \mathcal{Z}$ we have*

$$|g(z_1, \ldots, z_n) - g(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n)| \le c.$$

Let $Z_1, \ldots, Z_n$ be n independent random variables taking values in $\mathcal{Z}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$

$$|g(Z_1, \ldots, Z_n) - \mathbb{E} g(Z_1, \ldots, Z_n)| \le c\sqrt{\frac{n}{2}\log(2/\delta)}$$

Error Bound with Rademacher Complexity

Let $z_i = (x_i, y_i)$ and

$$g(z_1, \ldots, z_n) = \sup_{f \in \mathcal{H}} \left[\mathcal{E}(f) - \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) \right].$$

Assume[2] $|\ell(y', y)| \leq c$. We recall that for any two functions $\alpha, \beta : \mathcal{X} \to \mathbb{R}$, we have $\sup_x \alpha(x) - \sup_x \beta(x) \leq \sup_x |\alpha(x) - \beta(x)|$. Therefore

$$\begin{aligned} |g(z_1, \ldots, z_n) - g(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n)| \\ \leq \frac{1}{n} |\ell(f(x_i), y_i) - \ell(f(x_i'), y_i')| \\ \leq \frac{2c}{n} \end{aligned}$$

We can apply McDiarmid's inequality...

[2] This might require us to assume bounded inputs/outputs.

Error Bound with Rademacher Complexity

We have that for any $\delta > 0$, with probability at least $1 - \delta$,

$$\sup_{f \in \mathcal{H}} \mathcal{E}(f) - \mathcal{E}_n(f) \leq \mathbb{E}\left[\sup_{f \in \mathcal{H}} \mathcal{E}(f) - \mathcal{E}_n(f)\right] + c\sqrt{\frac{2\log(2/\delta}{n}}.$$

By applying our analysis in terms of the Rademacher complexity, we have also that

$$\sup_{f \in \mathcal{H}} \mathcal{E}(f) - \mathcal{E}_n(f) \leq 2L\,\mathcal{R}(\mathcal{H}) + c\sqrt{\frac{2\log(2/\delta)}{n}}.$$

Holds with probability at least $1 - \delta$

Recap

We have shown that the generalization error of an algorithm learning a function on a space of hypotheses $\mathcal{H}$ can be controlled in terms of the Rademacher complexity of such space...

Note. This applies to *any* algorithm, not just ERM![3]

But in general... when is the Rademacher complexity of $\mathcal{H}$ finite? And not too large?

We started from the observation that finite spaces were not that good for our purposes... So let's consider some other spaces!

[3] of course this would leave an outstanding term $\mathcal{E}_n(f_n) - \mathcal{E}_n(f_*)$... but this is a question for another day!

Rademacher Complexity In Practice...

Caveats in using Rademacher Complexity

With Rademacher complexity we now have a tool to study the theoretical properties of the ERM estimator (possibly others)

$$f_S = \operatorname*{arg\,min}_{f \in \mathcal{H}} \; \mathcal{E}_S(f)$$

Caveat:

Caveats in using Rademacher Complexity

With Rademacher complexity we now have a tool to study the theoretical properties of the ERM estimator (possibly others)

$$f_S = \arg\min_{f \in \mathcal{H}} \ \mathcal{E}_S(f)$$

Caveat: we need $\mathcal{R}(\mathcal{H})$ to be finite!

This opens two main questions:

- For which spaces can we "control" $\mathcal{R}(\mathcal{H})$?
- How to solve such constrained optimization problem?

Example - Linear Spaces

Let $\mathcal{X} = \mathbb{R}^d$ and consider a space of linear hypotheses

$$\mathcal{H} = \left\{ f \mid f(x) = \langle x, w \rangle\,,\ \forall x \in \mathcal{X},\ \exists w \in \mathbb{R}^d \right\}.$$

We want to study the Rademacher complexity of $\mathcal{H}$.

$$\begin{aligned}
\mathcal{R}_n(\mathcal{H}) &= \mathbb{E} \sup_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i) \\
&= \frac{1}{n} \mathbb{E} \sup_{f \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i \langle x_i, w_i \rangle \\
&= \frac{1}{n} \mathbb{E} \sup_{f \in \mathcal{H}} \left\langle \sum_{i=1}^{n} \sigma_i x_i, w_i \right\rangle \\
&\leq \frac{1}{n} \mathbb{E} \| \sum_{i=1}^{n} \sigma_i x_i \| \underbrace{\sup_{w \in \mathbb{R}^d} \|w\|}_{+\infty!}
\end{aligned}$$

Obtained applying Cauchy-Schwartz $\langle x, w \rangle \leq \|x\| \|w\|$.

Example - Balls in Linear Spaces

Let us restrict ourselves to balls in $\mathcal{H}$

$$\mathcal{H}_\gamma = \left\{ f \mid f(x) = \langle x, w \rangle \,,\ \forall x \in \mathcal{X},\ \exists w \in \mathbb{R}^d, \|w\| \leq \gamma \right\}.$$

Then,

$$\mathcal{R}_n(\mathcal{H}_\gamma) \leq \frac{\gamma}{n} \mathbb{E} \| \sum_{i=1}^{n} \sigma_i x_i \|$$

By noting that $\| \sum_{i=1}^n \sigma_i x_i \| = \left(\| \sum_{i=1}^n \sigma_i x_i \|^2 \right)^{1/2}$ and applying Jensen's inequality (Or simply the concavity of the square root), we have

$$\mathbb{E} \| \sum_{i=1}^{n} \sigma_i x_i \| \leq \left(\mathbb{E} \| \sum_{i=1}^{n} \sigma_i x_i \|^2 \right)^{1/2}$$

Example - Balls in Linear Spaces (Cont.)

Now

$$\mathbb{E}\|\sum_{i=1}^{n}\sigma_i x_i\|^2 = \mathbb{E}\sum_{i,j=1}^{n}\sigma_i\sigma_j\left\langle x_i, x_j\right\rangle$$

$$= \mathbb{E}_S\left[\sum_{i,j\neq 1}^{n}\mathbb{E}_\sigma[\sigma_i\sigma_j]\left\langle x_i, x_j\right\rangle + \sum_{i=1}^{n}\mathbb{E}_\sigma[\sigma_i^2]\|x_i\|^2\right]$$

Since the σ_i are independent and have zero mean, we have $\mathbb{E}_\sigma[\sigma_i\sigma_j] = 0$ for $i \neq j$ and $\mathbb{E}_\sigma[\sigma_i^2] = 1$. Therefore

$$\mathbb{E}\|\sum_{i=1}^{n}\sigma_i x_i\|^2 \leq \mathbb{E}_S\sum_{i=1}^{n}\|x_i\|^2$$

Example - Balls in Linear Spaces (Cont.)

Therefore, if we assume the input points to be bounded as well (e.g. in a ball of radius B in $\mathbb{R}^d$), we have

$$\begin{aligned}\mathcal{R}_n(\mathcal{H}) &\leq \frac{\gamma}{n}\left(\mathbb{E}_S \sum_{i=1}^{n} \|x_i\|^2\right)^{1/2} \\ &\leq \frac{\gamma}{n}\sqrt{nB^2} \\ &= \frac{\gamma B}{\sqrt{n}}\end{aligned}$$

Note. As expected, we have a bound on the generalization error that:

- decreases as n increases, but that becomes more and,
- becomes less meaningful as γ increases (since we are giving too much “freedom” to our learning algorithm to choose a function).

Example - Reproducing Kernel Hilbert Spaces

Following the example of spaces of linear hypotheses, we can think of generalizing the result also to RKHS...

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a bounded kernel, namely $k(x, x) \leq \kappa^2$ for any $x \in \mathcal{X}$ (e.g. $\kappa = 1$ for Gaussian or Abel kernels).

Let $\mathcal{H}$ be the RKHS associated to k and $\mathcal{H}_\gamma$ the space of $f \in \mathcal{H}$ such that $\|f\|_{\mathcal{H}} \leq \gamma$.

Then, we only need to replace each x_i with $k(x_i, \cdot)$ in our analysis for linear hypotheses and obtain

$$\mathcal{R}(\mathcal{H}_\gamma) \leq \frac{\gamma\kappa}{\sqrt{n}}$$

Constrained Optimization

The examples above show that considering the optimization over the entire space $\mathcal{H}$ is not a good idea (at least for Rademacher complexity)...

...But so far we have mostly seen examples of this form!

$$w_{S,\lambda} = \underset{w \in \mathbb{R}^d}{\arg\min} \; \mathcal{E}_S(w) + \lambda \|w\|^2$$

Does it mean that we cannot study the theoretical properties of Tikhonov regularization?

Constrained Optimization

The examples above show that considering the optimization over the entire space $\mathcal{H}$ is not a good idea (at least for Rademacher complexity)...

...But so far we have mostly seen examples of this form!

$$w_{S,\lambda} = \underset{w \in \mathbb{R}^d}{\arg\min}\ \mathcal{E}_S(w) + \lambda \|w\|^2$$

Does it mean that we cannot study the theoretical properties of Tikhonov regularization?

well... yes and no.

Rademacher and Tikhonov

Note. while it's true that Tikhonov considers all $w \in \mathbb{R}^d$, it does not *need to*...

Since $w_{S,\lambda}$ is the minimizer of the regularized problem, we have

$$\mathcal{E}_S(w_{S,\lambda}) + \lambda \left\| w_{S,\lambda} \right\|^2 \leq \mathcal{E}_S(0) + \lambda \left\| 0 \right\|^2$$

Assume for simplicity $\ell(y, y') \leq M^2$ for a constant $M > 0$, then

$$\left\| w_{S,\lambda} \right\| \leq \sqrt{\frac{\mathcal{E}_S(0)}{\lambda}} \leq \frac{M}{\sqrt{\lambda}}$$

namely we can restrict Tikhonov to $\mathcal{H}_\gamma$ with $\gamma = \frac{M}{\sqrt{\lambda}}$

Rademacher and Tikhonov (II)

Then, assuming $w_* \in \mathcal{H}$, we can consider the following decomposition of the excess risk

$$\mathbb{E}[\mathcal{E}(w_{S,\lambda}) - \mathcal{E}(w_*)] =$$

$$\underbrace{\mathbb{E}[\mathcal{E}(w_{S,\lambda}) - \mathcal{E}_S(w_{S,\lambda})]}_{\substack{\text{Rademacher} \\ \leq \frac{2LMB}{\sqrt{n\lambda}}}} + \underbrace{\mathbb{E}[\mathcal{E}_S(w_{S,\lambda}) - \mathcal{E}_S(w_*)]}_{\leq ?} + \underbrace{\mathbb{E}[\mathcal{E}_S(w_*) - \mathcal{E}(w_*)]}_{=0}$$

Rademacher and Tikhonov (III)

We can bound the remaining term by adding $\lambda \left\| w_{S,\lambda} \right\|^2$ and adding and removing $\lambda \left\| w_* \right\|^2$

$$
\begin{aligned}
\mathcal{E}_S(w_{S,\lambda}) - \mathcal{E}_S(w_*) \\
&\leq (\mathcal{E}_S(w_{S,\lambda}) + \lambda \left\| w_{S,\lambda} \right\|^2 - (\mathcal{E}_S(w_*) + \lambda \left\| w_* \right\|^2) + \lambda \left\| w_* \right\|^2 \\
&\leq \lambda \left\| w_* \right\|^2
\end{aligned}
$$

Rademacher and Tikhonov (Conclusion)

Putting everything together we conclude that

$$\mathbb{E}[\mathcal{E}(w_{S,\lambda}) - \mathcal{E}(w_*)] \leq \frac{2LMB}{\sqrt{n\lambda}} + \lambda \left\| w_* \right\|^2$$

Choosing $\lambda(n)$ to minimize this upper bound yields

$$\lambda(n) = \frac{(LMB)^{2/3}}{\left\| w_* \right\|^{4/3} n^{1/3}}$$

And an overall rate of

$$\mathbb{E}[\mathcal{E}(w_{S,\lambda(n)}) - \mathcal{E}(w_*)] \leq \frac{3(LMB)^{2/3} \left\| w_* \right\|^{2/3}}{n^{1/3}}$$

Ivanov Regularization

This is odd... from our analysis of Rademacher, if we took $\gamma = \|w_*\|$ and solved the so-called Ivanov regularization problem

$$w_{S,\gamma} = \arg\min_{\|w\| \leq \gamma} \mathcal{E}_S(w)$$

we would have a much faster excess risk bound

$$\mathbb{E}[\mathcal{E}(w_{S,\gamma}) - \mathcal{E}(w_*)] \leq O(\frac{1}{\sqrt{n}})$$

This is mainly because Rademacher complexity is not suited to study Tikhonov regularization...

...however, the observation above makes the Ivanov regularization a good strategy to obtain a predictor

How can we obtain $w_{S,\gamma}$ in practice?

Projected Gradient Descent

When $F : \mathbb{R}^d \to \mathbb{R}$ is a smooth convex function and $C \subset \mathbb{R}^d$ is a convex set, we can solve the constrained optimization

$$\min_{w \in C} F(w)$$

with a variant of GD: Projected Gradient Descent (PGD). Let

$$\Pi_C(w) = \arg\min_{z \in C} \|z - w\|^2$$

the projection of w onto C. Then, starting from w_0, PGD produces the sequence $(w_k)_{k \in \mathbb{N}}$ such that

$$w_{k+1} = \Pi_C(w_k - \eta \nabla F(w_k))$$

PGD on Euclidean Balls

Let's go back to the Ivanov regularization problem

$$w_{S,\gamma} = \underset{\|w\| \leq \gamma}{\arg\min}\ \mathcal{E}_S(w)$$

This corresponds to the constrained optimization problem with $F(\cdot) = \mathcal{E}_S(\cdot)$ and $C = \mathcal{H}_\gamma$ the ball of radius γ.

Given $w \in \mathcal{H} = \mathbb{R}^d$, projecting to the ball of radius γ yields

$$\Pi_{\mathcal{H}_\gamma}(w) = \begin{cases} w & \text{if } \|w\| \leq \gamma \\ \frac{\gamma}{\|w\|} w & \text{otherwise} \end{cases}$$

Therefore PGD for Ivanov regularization on Eucliden balls is as efficient as GD on the entire space!

...and what about convergence rates?

Convergence of PGD

Theorem (PGD Rates). Let F be convex and M-smooth. Assume F admits a minimum in $w_* \in C \subseteq R^d$ closed and convex set. Let $(w_k)_{k=1}^K$ be a sequence produced by PGD with $\eta = 1/M$. Then

$$F(w_K) - F(w_*) \leq \frac{M}{2K} \|w_0 - w_*\|^2$$

Proof

Lemma. *Let $z \in \mathbb{R}^d$ then for any $y \in C$*

$$(z - \Pi_C(z))^\top (y - \Pi_C(z)) \leq 0$$

Now, take $z = w - \frac{1}{M}\nabla F(w)$ and $w' = \Pi_C(z)$ the PGD step. Applying the Lemma yields

$$(w - w')^\top (y - w') \leq \frac{1}{M}\nabla F(w)^\top (y - w')$$

or equivalently

$$-M(w' - w)^\top (w' - y) \geq \nabla F(w)^\top (w' - y)$$

Proof

Proposition. *For any $y \in C$*

$$F(w') \leq F(y) + M(w' - w)^\top (y - w) - \frac{M}{2} \|w' - w\|^2$$

Proof.

$$\begin{aligned}
F(w') - F(y) &= F(w') - F(w) + F(w) - F(y) \\
&\leq \nabla F(w)^\top (w' - w) + \frac{M}{2} \|w' - w\|^2 + \nabla F(w)^\top (w - y) \\
&= \nabla F(w)^\top (w' - y) + \frac{M}{2} \|w' - w\|^2 \\
&\leq -M(w' - w)^\top (w' - y) + \frac{M}{2} \|w' - w\|^2
\end{aligned}$$

Adding and removing w inside $(w' - w)$ yields

$$F(w') - F(y) \leq -M(w' - w)^\top (w - y) - \frac{M}{2} \|w' - w\|^2$$

as required.

Proof

The term $M(w' - w)$ now plays the same role originally played by $\nabla F(w)$ in the proof of GD. Consider

$$F(w_{k+1}) - F(w_*) \leq M(w_{k+1} - w_k)^\top (w_k - w_*) - \frac{M}{2} \|w_{k+1} - w_k\|^2$$

Then, by adding and removing $\frac{M}{2} \|w_k - w_*\|$ and "completing the square", we obtain

$$F(w_{k+1}) - F(w_*) \leq \frac{M}{2}(\|w_k - w_*\|^2 - \|w_{k+1} - w_*\|^2)$$

Proof

Exploiting the telescopic sum

$$\sum_{k=1}^{K}(F(w_{k+1}) - F(w_*)) \leq \frac{M}{2}\sum_{k=1}^{K}(\|w_k - w_*\|^2 - \|w_{k+1} - w_*\|^2)$$
$$\leq \frac{M}{2}\|w_0 - w_*\|^2$$

and the fact that the PGD algorithm is decreasing[4], yields the required result.

[4] **Exercise.** Why?

Wrapping Up

- Unsatisfied by being able to control the generalization error of a learning algorithm only when considering finite spaces of hypotheses, we payed more careful attention to the way we bounded it.
- We observed that by looking at the worst generalization error in a class of functions (rather than the sum of all such errors which might be too large), can be controlled in terms of the *Rademacher complexity* of such space of hypotheses.
- We concluded showing that for the case of spaces of linear hypotheses, or more generally for balls in a RKHS, such complexity is bounded by a finite quantity that depends on the number of training points and the radius of the ball.
- We provided an efficient algorithm to solve the corresponding (constrained) ERM problem.