3.12 Norm-constrained hypothesis classes

[Theorem 11] Rademacher complexity of L_2 ball.

Let $\mathcal{F} = \{ z \mapsto w \cdot z : \|w\|_2 \leq B_2 \}$ bounds on weight vectors.

Assume $\mathbb{E}_{z \sim p^*}[\|Z\|_2^2] \leq C_2^2$, Then

$$R_n(\mathcal{F}) \leq \frac{B_2 C_2}{\sqrt{n}}$$

Pf:

$$\begin{aligned}
R_n(\mathcal{F}) &= \frac{1}{n} \mathbb{E}\Big[\sup_{\|w\|_2 \leq B_2} \sum_{i=1}^n \sigma_i (w \cdot Z_i) \Big] \\
&\leq \frac{1}{n} \mathbb{E}\Big[\sup_{\|w\|_2 \leq B_2} \|w\|_2 \Big\| \sum_{i=1}^n \sigma_i Z_i \Big\|_2 \Big] \quad (\text{Hölder ineq.}) \\
&\leq \frac{B_2}{n} \mathbb{E}\Big[\Big\| \sum_{i=1}^n \sigma_i Z_i \Big\|_2 \Big] \\
&\leq \frac{B_2}{n} \sqrt{ \mathbb{E}\Big[\Big\| \sum_{i=1}^n \sigma_i Z_i \Big\|_2^2 \Big] } \quad (\text{concavity of sqrt}) \\
&= \frac{B_2}{n} \sqrt{ \mathbb{E}\Big[\sum_{i=1}^n \|\sigma_i Z_i\|_2^2 \Big] } \quad \text{expectation of cross term} = 0. \\
&= \frac{B_2}{n} \sqrt{ \mathbb{E}\Big[\sum_{i=1}^n \|Z_i\|_2^2 } \\
&\leq \frac{B_2 C_2}{\sqrt{n}}. \qquad \square
\end{aligned}$$

[Theorem 12] Rademacher complexity of L_1 ball

Assume $\|Z_i\|_\infty \leq C_\infty$ with prob. 1 for all data points $i = 1, \cdots, n$.

Then $$R_n(\mathcal{F}) \leq \frac{B_1 C_\infty \sqrt{2 \log(2d)}}{\sqrt{n}}$$

Pf: key : L_1 ball is the convex hull of the following

$$W = \bigcup_{j=1}^{d} \{ B_1 e_j, -B_1 e_j \}$$

$$\Rightarrow R_n(\mathcal{F}) = \mathbb{E}\Big[\sup_{w \in W} \frac{1}{n} \sum_{i=1}^n \sigma_i (w \cdot Z_i) \Big]$$

We have $w \cdot Z_i \leq \|w\|_1 \|Z_i\|_\infty \leq B_1 C_\infty$. By Massart's finite lemma

$$R_n(\mathcal{F}) \leq \sqrt{\frac{2M^2 \log|\mathcal{F}|}{n}} = \sqrt{\frac{2 B_1^2 C_\infty^2 \log(2d)}{n}} \qquad \square$$

Recall: (i) p-norm decrease : $\|w\|_p \geq \|w\|_q$, $p \leq q$

(ii) ball size increase : $\{ w : \|w\|_p \leq B \} \subseteq \{ w : \|w\|_q \leq B \}$, $p \leq q$

(iii) Ball size increase: $\{w : \|w\|_p \le B\} \subseteq \{w : \|w\|_q \le B\}$, $p \le q$

Ramifications under sparsity:

(i) L_1 regularization is often used when we believe most features are irrelevant — $s \ll d$ non-zero entries.

(ii) Assume $\|w\|_\infty \le 1$, $\|x\|_\infty \le 1$. It's sufficient to consider $\|w\|_1 \le B_1 = s$. Then $R_n(H) = O\left(\frac{s\sqrt{\log d}}{\sqrt{n}}\right)$, which means the number of relevant features (s) controls the complexity.

(iii) In contrast, if we use L_2 regularization, we would have $B_2 = \sqrt{s}$, $C_2 = \sqrt{d}$, $R_n(H) = O\left(\frac{s\sqrt{d/s}}{\sqrt{n}}\right)$.

If $s \ll d$, L_1 better. If $s = d$, L_2 better.

From hypothesis class to loss class (for binary classification)

- Since now we talk about hypothesis class containing real-valued functions, the previouly argument for $s(n, H) = s(n, A)$ doesn't hold.
- Consider ϕ for binary classification that only depends on the margin $m = yx \cdot w$. For example:
  - Zero-one loss: $\phi(m) = \mathbb{1}\{m \le 0\}$
  - Hinge loss: $\phi(m) = \max\{0, 1-m\}$
- Let $w \in W$ be a set of weight vectors. The loss class corresponding to these weight vectors is then:

$$A = \{(x, y) \to \phi(yx \cdot w) : w \in W\}$$

Simply think our data points as $z = xy$

$$F = \{z \mapsto wz : w \in W\}.$$

Therefore, we can rewrite:

$$A = \phi \circ F$$

where we could apply the composition rule for Lipschitz functions.

Since $\mathbb{1}$ is not Lipschitz, we make some weaker statement.

First we define a margin-sensitive version of zero-one loss:

$$\phi_\gamma(m) = \mathbb{1}\{m \le \gamma\}$$

and $\quad L_\gamma(m) = E_{z\sim p^*}[\phi_\gamma(w\cdot z)]$, the associated expected risk

---

[Theorem 13] margin-sensitive zero-one loss for linear classifiers.

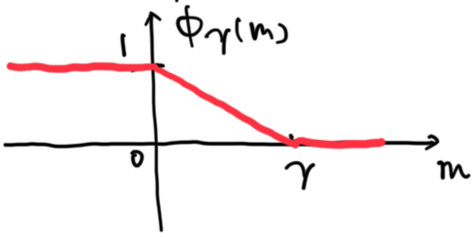
Let F be a set of linear functions. L_γ is defined above.

Let $\hat{w}$ and w^* be the weight vectors associated with the empirical and expected risk minimizers w.r.t. $\tilde{\phi}_\gamma$. With prob. $\geq 1-\delta$,

$$L_0(\hat{w}) \le L_\gamma(w^*) + \frac{4R_n(F)}{\gamma} + \sqrt{\frac{2\log(2/\delta)}{n}}$$

Pf: Problem: ϕ_γ is not Lipschitz for any γ.

Define $\quad \tilde{\phi}_\gamma(m) = \min\{1, \max\{0, 1-\frac{m}{\gamma}\}\}$



Then the Rademacher complexity of this intermediate loss class can be bounded:

$$R_n(\tilde{\phi}_\gamma \circ F) = \frac{R_n(F)}{\gamma}$$

By Theorem 9:

$$\tilde{L}_\gamma(\hat{w}) \le \hat{\tilde{L}}_\gamma(w^*) + \frac{4R_n(F)}{\gamma} + \sqrt{\frac{2\log(2/\delta)}{n}}$$

Note that $\phi_0 \le \tilde{\phi}_\gamma \le \phi_\gamma$, so

$$L_0(w) \le \tilde{L}_\gamma(w) \le L_\gamma(w)$$

$$\Rightarrow \hat{L}_0(\hat{w}) \le L_\gamma(w^*) + \frac{4R_n(F)}{\gamma} + \sqrt{\frac{2\log(2/\delta)}{n}} \qquad \square$$