Chapter 1
Risk Minimization and Rademacher Complexity

Ernest K. Ryu
Seoul National University

Mathematical Machine Learning Theory
Spring 2024

Outline

Supervised learning setup

Given data $X_1, \ldots, X_N \in \mathcal{X}$ and corresponding labels $Y_1, \ldots, Y_N \in \mathcal{Y}$, where $\mathcal{X}$ is the data space $\mathcal{Y}$ is the label space. Goal is to learn a function $f\colon \mathcal{X} \to \mathcal{Y}$ such that $f(X) \approx Y$ for new data-label pairs (X, Y).

More formally, let $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a *loss function* that quantifies the size of the error. Often, $\ell(y', y) \geq 0$ for all $y', y \in \mathcal{Y}$. Assume $(X_i, Y_i) \overset{\mathrm{IID}}{\sim} P$. We further formalize the goal as

$$\underset{f}{\text{minimize}} \quad \mathop{\mathbb{E}}_{(X,Y)\sim P}[\ell(f(X), Y)].$$

For now, consider the minimization over all functions f, although we will soon see that we must restrict the class of functions.

Precisely speaking the expectation is well defined only for appropriately measurable functions ℓ and f. In this course, we will not seriously engage with the issue of measurability, but I will point out the issue when relevant.

Supervised learning setup

Sometimes, actually, we don't want the "prediction" of f to be exactly the same type as the label $Y \in \mathcal{Y}$.

Assume $(X_i, Y_i) \stackrel{\text{IID}}{\sim} P$. More generally, let $f \colon \mathcal{X} \to \tilde{\mathcal{Y}}$ and $\ell \colon \tilde{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$. We formalize the goal as

$$\underset{f \in \mathcal{F}}{\text{minimize}} \quad \mathbb{E}_{(X,Y)\sim P}\left[\ell(f(X), Y)\right] \quad .$$

Example) K-class classification with cross-entropy loss, where $\mathcal{Y} = \{1, 2, \dots, K\}$ and

$$\tilde{\mathcal{Y}} = \Delta_K = \{(p_1, \dots, p_K \,|\, p_1, \dots, p_K \geq 0,\, p_1 + \cdots + p_K = 1\}.$$

I.e., label Y is a single class, but the prediction is a probability distribution over the K classes. The *cross-entropy* loss is

$$\ell^{\mathrm{CE}}(y', y) = -\log\left(\frac{\exp(y'_y)}{\sum_{k=1}^{K} \exp(y'_k)}\right) > 0.$$

Expected risk

The *expected risk*, also called the *true risk*, is

$$\mathcal{R}[f] = \mathop{\mathbb{E}}_{(X,Y)\sim P}[\ell(f(X), Y)].$$

Our goal is to solve

$$\underset{f}{\text{minimize}} \quad \mathcal{R}[f].$$

We call

$$\mathcal{R}^\star = \inf_f \mathcal{R}[f]$$

the *Bayes risk* or the *optimal risk*, where the infimum is over all functions.

Bayes predictor

Optimal $f^\star\colon \mathcal{X} \to \tilde{\mathcal{Y}}$ attaining the Bayes risk is characterized as follows.

By the law of iterated expectations, we have

$$\begin{aligned}\mathcal{R}[f] &= \underset{(X,Y)\sim P}{\mathbb{E}}[\ell(f(X), Y)] \\ &= \underset{X\sim P_X}{\mathbb{E}}\left[\underset{Y\sim P_{Y\mid X}}{\mathbb{E}}[\ell(f(X), Y) \mid X]\right].\end{aligned}$$

Then, the *Bayes predictor* $f^\star$, defined by

$$f^\star(X) \in \underset{y' \in \tilde{\mathcal{Y}}}{\operatorname{argmin}} \underset{Y\sim P_{Y\mid X}}{\mathbb{E}}[\ell(y', Y) \mid X],$$

attains the Bayes risk, i.e.,

$$\mathcal{R}^\star = \mathcal{R}[f^\star].$$

(So, the Bayes predictor is the exact/perfect solution to given ML task.)

Theorem

Let $f^\star$ be such that

$$f^\star(X) \in \operatorname*{argmin}_{y' \in \tilde{\mathcal{Y}}} \mathop{\mathbb{E}}_{Y \sim P_{Y|X}} [\ell(y', Y) \mid X] \qquad \forall X \in \mathcal{X}.$$

Then,

$$\mathcal{R}[f] \geq \mathcal{R}[f^\star] \qquad \forall\, f.$$

(We do not know whether $f^\star$ exists or whether it is unique.)

Proof. Since

$$\mathop{\mathbb{E}}_{Y \sim P_{Y|X}} [\ell(f(X), Y) \mid X] \geq \mathop{\mathbb{E}}_{Y \sim P_{Y|X}} [\ell(f^\star(X), Y) \mid X] \qquad \forall\, X \in \mathcal{X},$$

by the law of iterated expectations, we have

$$\begin{aligned}
\mathcal{R}[f] &= \mathop{\mathbb{E}}_{X \sim P_X} \left[\mathop{\mathbb{E}}_{Y \sim P_{Y|X}} [\ell(f(X), Y) \mid X] \right] \\
&\geq \mathop{\mathbb{E}}_{X \sim P_X} \left[\mathop{\mathbb{E}}_{Y \sim P_{Y|X}} [\ell(f^\star(X), Y) \mid X] \right] = \mathcal{R}[f^\star].
\end{aligned}$$

$\square$

Example: Binary classification

Consider $\tilde{\mathcal{Y}} = \mathcal{Y} = \{-1, +1\}$ and $\ell(y', y) = \mathbf{1}_{\{y' \neq y\}}$. So

$$\mathcal{R}[f] = \mathop{\mathbb{E}}_{(X,Y)\sim P}[\ell(f(X), Y)] = \mathop{\mathbb{P}}_{(X,Y)\sim P}(f(X) \neq Y).$$

Then,

$$f^\star(X) = \begin{cases} -1 & \text{if } \mathbb{P}(Y = -1 \,|\, X) \geq \mathbb{P}(Y = +1 \,|\, X) \\ +1 & \text{if } \mathbb{P}(Y = +1 \,|\, X) < \mathbb{P}(Y = -1 \,|\, X) \end{cases}$$

(with ties broken arbitrarily) is a Bayes predictor, and

$$\mathcal{R}^\star = \mathop{\mathbb{E}}_{X\sim P_X}[\min\{\mathbb{P}(Y = -1 \,|\, X), \mathbb{P}(Y = +1 \,|\, X)\}].$$

Example: Regression with squared loss

Consider $\tilde{\mathcal{Y}} = \mathcal{Y} = \mathbb{R}$ and $\ell(y', y) = (y' - y)^2$. Then

$$\begin{aligned} f^\star(X) &= \operatorname*{argmin}_{y' \in \mathbb{R}} \mathop{\mathbb{E}}_{Y \sim P_{Y \mid X}} [(y' - Y)^2 \mid X] \\ &= \operatorname*{argmin}_{y' \in \mathbb{R}} \mathop{\mathbb{E}}_{Y \sim P_{Y \mid X}} [(y' - \mathbb{E}[Y \mid X])^2 + (\mathbb{E}[Y \mid X] - Y)^2 \\ &\qquad\qquad\qquad\qquad + 2(y' - \mathbb{E}[Y \mid X])(\mathbb{E}[Y \mid X] - Y) \mid X] \\ &= \operatorname*{argmin}_{y' \in \mathbb{R}} \mathop{\mathbb{E}}_{Y \sim P_{Y \mid X}} [(y' - \mathbb{E}[Y \mid X])^2 + (\mathbb{E}[Y \mid X] - Y)^2 \mid X] \\ &= \mathbb{E}[Y \mid X]. \end{aligned}$$

Note that only the blue term depends on y'.

So the conditional mean $\mathbb{E}[Y \mid X]$ is the optimal Bayes predictor, and

$$\mathcal{R}^\star = \mathop{\mathbb{E}}_{X \sim P_X} [\mathrm{Var}(Y \mid X)]$$

is the expected conditional variance of Y.

Excess risk and empirical risk

Think of $\mathcal{R}^\star$ as the optimal (smallest) risk one could achieve, in principle, with infinite data and compute.

Define *excess risk* as

$$\mathcal{R}[f] - \mathcal{R}^\star,$$

which is the risk f achieve compared to the baseline of $\mathcal{R}^\star$.
In practice, we do not have access to the true risk. We instead have access to the *empirical risk*

$$\hat{\mathcal{R}}[f] = \frac{1}{N}\sum_{i=1}^{N} \ell(f(X_i), Y_i).$$

However,

$$\underset{f}{\text{minimize}} \quad \hat{\mathcal{R}}[f],$$

where the minimization is over all functions, is a bad idea as it leads to severe overfitting.

Function class (hypothesis set)

We write $\mathcal{F}$ to denote a *function class* (also called a *hypothesis set*) used in an ML algorithm.

$\mathcal{F}$ is a "small" subset of functions; it is not all functions.

- Considering all functions would be computationally expensive.
- Having a "large" function class $\mathcal{F}$ causes overfitting (large estimation error, large Rademacher complexity), as we discuss soon.

$\mathcal{F}$ is often not a vector space.

- We often impose compactness, and $\mathcal{F}$ becomes a sub*set* of a vector space.
- In deep learning, neural networks depend on their parameters nonlinearly, and $\mathcal{F}$ becomes a "manifold" within a larger function (vector) space.

Empirical risk minimization

Eempirical risk minimization considers

$$\hat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} \hat{\mathcal{R}}[f]$$

or

$$\hat{f} \approx \operatorname*{argmin}_{f \in \mathcal{F}} \hat{\mathcal{R}}[f].$$

We use the notation $X \approx \operatorname{argmin}$ to say that X is an approximate minimizer. The consequence of solving the minimization inexactly will be addressed later when we discuss optimization error.

Risk decomposition

Let $\hat{f}$ be the output of an ML algorithm. (Usually approximate empirical risk minimization over a parameterized class of functions.)

Our analyses will be based on the *risk decomposition*:

$$\mathcal{R}[\hat{f}] - \mathcal{R}^\star = \underbrace{(\mathcal{R}[\hat{f}] - \inf_{f' \in \mathcal{F}} \mathcal{R}[f'])}_{=\text{Estimation error} \geq 0} + \underbrace{(\inf_{f' \in \mathcal{F}} \mathcal{R}[f'] - \mathcal{R}^\star)}_{=\text{Approximation error} \geq 0}$$

Approximation error only depends on $\mathcal{F}$, P, and ℓ; it does not depend on the data or the choice of ML algorithm. If $\mathcal{F}$ is sufficiently expressive, i.e., if $\mathcal{F}$ can approximate the optimal Bayes predictor $f^\star$ well, then the approximation error will be small.

Estimation error depends on $\hat{f}$, which, in turn, depends on the data $\{(X_i, Y_i)\}_{i=1}^N$ and the ML algorithm.

Risk decomposition

Goal is to show excess risk is small, i.e.,

$$\mathcal{R}[\hat{f}] - \mathcal{R}^\star \leq \text{small},$$

by showing

$$\text{Estimation error} = \mathcal{R}[\hat{f}] - \inf_{f' \in \mathcal{F}} \mathcal{R}[f'] \leq \text{small}$$

and

$$\text{Approximation error} = \inf_{f' \in \mathcal{F}} \mathcal{R}[f'] - \mathcal{R}^\star \leq \text{small}.$$

Note, estimation error is random (because $\hat{f}$ is random), and approximation error is deterministic.

To argue that the excess risk is "small", we need to show that estimation error is either small in expectation or small with high probability.

Bias-variance tradeoff

Goal is to show excess risk is small, i.e.,

$$\mathcal{R}[\hat{f}] - \mathcal{R}^\star \leq \text{small}$$

by showing

$$\text{Estimation error} = \mathcal{R}[\hat{f}] - \inf_{f' \in \mathcal{F}} \mathcal{R}[f'] \leq \text{small}$$

and

$$\text{Approximation error} = \inf_{f' \in \mathcal{F}} \mathcal{R}[f'] - \mathcal{R}^\star \leq \text{small}.$$

Typically, estimation error goes down as N goes up, but it goes up as $\mathcal{F}$ becomes large.

Typically, approximation error goes down to 0 as $\mathcal{F}$ becomes large.
(By universal approximation theorems.)

Bias-variance tradeoff

In most cases, large N is better,[1] but large $\mathcal{F}$ is not always better, even though processing large $\mathcal{F}$ requires more compute.

In traditional statistics and ML theory,[2] the best $\mathcal{F}$ is the solution of the *bias-variance tradeoff*, a trade-off between underfitting and overfitting.

Underfitting is loosely defined by the following conditions:

- high bias, low variance
- small estimation error, large approximation error
- small $\mathcal{F}$

Overfitting is loosely defined by the following conditions:

- low bias, high variance
- large estimation error, small approximation error
- large $\mathcal{F}$

[1] There are some counterintuitive counterexamples to this:
P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, Deep double descent: Where bigger models and more data hurt, *ICLR*, 2020.

[2] "Double-descent" and "benign overfitting" is the alternate modern view.

Universal approximation result

We will soon see why large $\mathcal{F}$ can increase estimation error.

However, typically, large $\mathcal{F}$ reduces approximation error

$$\text{Approximation error} = \inf_{f' \in \mathcal{F}} \mathcal{R}[f'] - \mathcal{R}^\star$$

due to *universal approximation theory*.

In this course, we won't get to this topic, but such results have the following flavor.

Theorem (Universal approximation theorem. Informal)

Let f_θ be an L-layer neural network with $L \geq 2$. If f_θ has sufficiently many neurons, then f_θ can approximate any function in the sense of L^p for any $p \in [1, \infty]$.

(It is possible to show a quantitative approximation result that describes the number of neurons needed to achieve an $\varepsilon > 0$ approximation.)

Corollary: If $\mathcal{F}$ large, neural network f_θ can approximate optimal Bayes predictor well, and approximation error ≈ 0.

Outline

Estimation error decomposition

$$
\begin{aligned}
\text{Estimation error} &= \mathcal{R}[\hat{f}] - \inf_{f' \in \mathcal{F}} \mathcal{R}[f'] \\
&= \mathcal{R}[\hat{f}] - \mathcal{R}[g] \qquad\qquad (\text{define } g = \operatorname*{argmin}_{f' \in \mathcal{F}} \mathcal{R}[f']) \\
&= (\mathcal{R}[\hat{f}] - \hat{\mathcal{R}}[\hat{f}]) + (\hat{\mathcal{R}}[g] - \mathcal{R}[g]) + (\hat{\mathcal{R}}[\hat{f}] - \hat{\mathcal{R}}[g]) \\
&\le \sup_{f \in \mathcal{F}} \{\mathcal{R}[f] - \hat{\mathcal{R}}[f]\} + \sup_{f \in \mathcal{F}} \{\hat{\mathcal{R}}[f] - \mathcal{R}[f]\} + (\hat{\mathcal{R}}[\hat{f}] - \hat{\mathcal{R}}[g]) \\
&\le \sup_{f \in \mathcal{F}} \{\mathcal{R}[f] - \hat{\mathcal{R}}[f]\} + \sup_{f \in \mathcal{F}} \{\hat{\mathcal{R}}[f] - \mathcal{R}[f]\} + \underbrace{(\hat{\mathcal{R}}[\hat{f}] - \inf_{f \in \mathcal{F}} \hat{\mathcal{R}}[f])}_{=\text{Optimization error} \approx 0}
\end{aligned}
$$

For now, assume opt. error is negligible. We'll bound opt. error later.

(This identity holds the same even if a minimizer g does not exist.)

Uniform bound

Ignoring the optimization error, we are left to bound

$$\sup_{f\in\mathcal{F}}\{\mathcal{R}[f]-\hat{\mathcal{R}}[f]\}+\sup_{f\in\mathcal{F}}\{\hat{\mathcal{R}}[f]-\mathcal{R}[f]\}$$

Sometimes, one proceeds with the

$$\sup_{f\in\mathcal{F}}\{\mathcal{R}[f]-\hat{\mathcal{R}}[f]\}+\sup_{f\in\mathcal{F}}\{\hat{\mathcal{R}}[f]-\mathcal{R}[f]\}\leq 2\sup_{f\in\mathcal{F}}\left|\mathcal{R}[f]-\hat{\mathcal{R}}[f]\right|,$$

and bound the RHS with a uniform bound on $\left|\mathcal{R}[f]-\hat{\mathcal{R}}[f]\right|$.

Why uniform convergence?

Loosely speaking, we will show

$$\sup_{f \in \mathcal{F}} \left| \mathcal{R}[f] - \hat{\mathcal{R}}[f] \right| \to 0,$$

i.e., show $\hat{\mathcal{R}} \overset{\text{uniform}}{\to} \mathcal{R}$, as $N \to \infty$. This is a standard argument.

This bound may seem pessimistic (loose), but it is crucial. Since $\hat{f} \approx \operatorname{argmin}_{f \in \mathcal{F}} \hat{\mathcal{R}}[f]$, the statistical dependence between $\hat{\mathcal{R}}$ and $\hat{f}$ is usually intractable.

By passing to the uniform bound, we eliminate $\hat{f}$ and thereby remove the statistical dependence between $\hat{\mathcal{R}}$ and $\hat{f}$. We now only need to deal with the randomness of $\hat{\mathcal{R}}$.

Expected error to PAC bound

Assume we can show

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\mathcal{R}[f]-\hat{\mathcal{R}}[f]\right|\right] < \text{small.}$$

Then we can show a concentration result

$$\sup_{f\in\mathcal{F}}\left|\mathcal{R}[f]-\hat{\mathcal{R}}[f]\right| < \varepsilon \qquad \text{with probability} > 1-\delta.$$

Using Markov, we can show

$$\sup_{f\in\mathcal{F}}\left|\mathcal{R}[f]-\hat{\mathcal{R}}[f]\right| < \frac{\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\mathcal{R}[f]-\hat{\mathcal{R}}[f]\right|\right]}{\delta} \qquad \text{w.p.} > 1-\delta.$$

However, we can obtain a much stronger bound with McDiarmid.

PAC bound with McDiarmid

Assume $0 \leq \ell(f(X), Y) \leq \ell_\infty$ for all $f \in \mathcal{F}$ and $(X, Y) \sim P$.[3]
Assumption holds if:

- 0-1 loss $\Phi_{0\text{-}1}$ is used; or
- Convex surrogate loss[4] is used, $f \in \mathcal{F}$ is continuous, $|\mathcal{F}| < |infty$, $|\mathcal{Y}| < \infty$, and $X \sim P$ has compact support (e.g. images with pixel values in $[0, 1]$).

Let $Z_i = (X_i, Y_i)$ for $i = 1, \ldots, N$, and let

$$H(Z_1, \ldots, Z_N) = \sup_{f \in \mathcal{F}} \left\{ \mathcal{R}[f] - \hat{\mathcal{R}}[f] \right\}$$

and use the McDiarmid inequality to obtain a PAC bound.

[3]So $0 \leq \ell(f(X), Y) \leq \ell_\infty$ for all $f \in \mathcal{F}$, P-almost surely.
[4]Convex functions are continuous.

PAC bound with McDiarmid

The bounded differences property

$$\left|H(\underbrace{Z_1,\ldots,Z_{i-1},Z_i,Z_{i+1},\ldots,Z_N}_{=\mathcal{D}}) - H(\underbrace{Z_1,\ldots,Z_{i-1},Z_i',Z_{i+1},\ldots,Z_N}_{=\mathcal{D}'})\right| \leq c$$

is the main condition to be checked.

To see this, note that

$$\hat{\mathcal{R}}[f](\mathcal{D}') - \hat{\mathcal{R}}[f](\mathcal{D}) = \frac{1}{N}\big(\ell(f(X_i'),Y_i') - \ell(f(X_i),Y_i)\big) \leq \frac{\ell_\infty}{N}.$$

Then we have

$$\begin{aligned}
&H(\mathcal{D}) - H(\mathcal{D}') \\
&= \sup_{f\in\mathcal{F}}\left\{\mathcal{R}[f] - \hat{\mathcal{R}}[f](\mathcal{D}') + \hat{\mathcal{R}}[f](\mathcal{D}') - \hat{\mathcal{R}}[f](\mathcal{D})\right\} - \sup_{f\in\mathcal{F}}\left\{\mathcal{R}[f] - \hat{\mathcal{R}}[f](\mathcal{D}')\right\} \\
&\leq \sup_{f\in\mathcal{F}}\left\{\mathcal{R}[f] - \hat{\mathcal{R}}[f](\mathcal{D}')\right\} + \sup_{f\in\mathcal{F}}\left\{\hat{\mathcal{R}}[f](\mathcal{D}') - \hat{\mathcal{R}}[f](\mathcal{D})\right\} - \sup_{f\in\mathcal{F}}\left\{\mathcal{R}[f] - \hat{\mathcal{R}}[f](\mathcal{D}')\right\} \\
&= \sup_{f\in\mathcal{F}}\left\{\hat{\mathcal{R}}[f](\mathcal{D}') - \hat{\mathcal{R}}[f](\mathcal{D})\right\} \leq \frac{\ell_\infty}{N}.
\end{aligned}$$

So $c = \frac{\ell_\infty}{N}$ and $|H(\mathcal{D}) - H(\mathcal{D}')| \leq \frac{\ell_\infty}{N}$ with a symmetric argument.

PAC bound with McDiarmid

Therefore, we conclude

$$\sup_{f\in\mathcal{F}} \left\{\mathcal{R}[f] \leq \hat{\mathcal{R}}[f]\right\} \leq \mathbb{E}\Big[\sup_{f\in\mathcal{F}} \left\{\mathcal{R}[f] - \hat{\mathcal{R}}[f]\right\}\Big] + \ell_\infty \sqrt{\frac{\log(1/\delta)}{2N}}$$

with probability $1-\delta$.

By the same reasoning, we have

$$\sup_{f\in\mathcal{F}} \left\{\hat{\mathcal{R}}[f] - \mathcal{R}[f]\right\} \leq \mathbb{E}\Big[\sup_{f\in\mathcal{F}} \left\{\hat{\mathcal{R}}[f] - \mathcal{R}[f]\right\}\Big] + \ell_\infty \sqrt{\frac{\log(1/\delta)}{2N}}$$

with probability $1-\delta$.

By a union bound, we have

$$\sup_{f\in\mathcal{F}} \left\{\hat{\mathcal{R}}[f] - \mathcal{R}[f]\right\} + \sup_{f\in\mathcal{F}} \left\{\mathcal{R}[f] - \hat{\mathcal{R}}[f]\right\}$$

$$\leq \mathbb{E}\Big[\sup_{f\in\mathcal{F}} \left\{\hat{\mathcal{R}}[f] - \mathcal{R}[f]\right\}\Big] + \mathbb{E}\Big[\sup_{f\in\mathcal{F}} \left\{\mathcal{R}[f] - \hat{\mathcal{R}}[f]\right\}\Big] + \ell_\infty \sqrt{\frac{2\log(2/\delta)}{N}}$$

with probability $1-\delta$.

Example: Finite number of models

We show examples of bounding the estimation error.

Consider $|\mathcal{F}| = m < \infty$, i.e., we are learning among a finite number of models. Let $\{f_1, \ldots, f_m\} = \mathcal{F}$ and

$$\hat{f} = \operatorname*{argmin}_{f_1,\ldots,f_m \in \mathcal{F}} \hat{\mathcal{R}}[f_i].$$

Assume $0 \le \ell(f(X), Y) \le \ell_\infty$ for all $f \in \mathcal{F}$ and $(X, Y) \sim P$. Since

$$\hat{\mathcal{R}}[f] - \mathcal{R}[f] = \frac{1}{N} \sum_{i=1}^{N} \underbrace{\ell(f(X_i), Y_i) - \mathbb{E}[\ell(f(X), Y)]}_{\text{zero-mean sub-Gauss. with } \tau^2 = \ell_\infty^2},$$

$\hat{\mathcal{R}}[f] - \mathcal{R}[f]$ is a zero-mean sub-Gaussian with $\tau^2 = \ell_\infty^2 / N$.

Then,

$$\mathbb{E}\Big[\sup_{f \in \mathcal{F}} \big\{\mathcal{R}[f] - \hat{\mathcal{R}}[f]\big\}\Big] \le \mathbb{E}\Big[\max_{i=1,\ldots,m} \big\{\hat{\mathcal{R}}[f_i] - \mathcal{R}[f_i]\big\}\Big]$$

$$\le \sqrt{\frac{2\ell_\infty^2}{N} \log m}.$$

Example: Finite number of models

Combining this with McDiarmid inequality,

$$\sup_{f\in\mathcal{F}} \left\{\mathcal{R}[f] - \hat{\mathcal{R}}[f]\right\} \le \sqrt{\frac{2\ell_\infty^2}{N}} \left(\sqrt{\log m} + \sqrt{\frac{\log(1/\delta)}{4}}\right)$$

with probability $1-\delta$. The same bound on $\sup_{f\in\mathcal{F}} \left\{\hat{\mathcal{R}}[f] - \mathcal{R}[f]\right\}$ can be obtained with the same argument.

Finally, we have

$$\begin{aligned}
\text{Estimation error} &= \mathcal{R}[\hat{f}] - \inf_{f'\in\mathcal{F}} \mathcal{R}[f'] \\
&\le \sup_{f\in\mathcal{F}} \left\{\mathcal{R}[f] - \hat{\mathcal{R}}[f]\right\} + \sup_{f\in\mathcal{F}} \left\{\hat{\mathcal{R}}[f] - \mathcal{R}[f]\right\} + \underbrace{\text{Opt. error}}_{=0} \\
&\le 2\sqrt{\frac{2\ell_\infty^2}{N}} \left(\sqrt{\log m} + \sqrt{\frac{\log(2/\delta)}{4}}\right)
\end{aligned}$$

with probability $1-\delta$.

ε-cover

We say $(\mathcal{F}, \|\cdot\|_\infty)$ is *totally bounded* if for any $\varepsilon > 0$, there is $m(\varepsilon) < \infty$ and $f_1, \ldots, f_{m(\varepsilon)} \in \mathcal{F}$ such that

$$\mathcal{F} \subseteq \bigcup_{i=1}^{m(\varepsilon)} \mathcal{B}(f_i, \varepsilon),$$

where $\mathcal{B}(f_i, \varepsilon) = \{f \in \mathcal{F} \,|\, \|f - f_i\|_\infty < \varepsilon\}$.

We say $f_1, \ldots, f_{m(\varepsilon)}$ is an ε-cover of size $m(\varepsilon)$.

(As an aside, in complete metric spaces, a set is compact if and only if it is closed and totally bounded.)

Example: Infinite models with covering number

Assume $\ell(\cdot, Y)$ is G-Lipschitz for all $Y \sim P_Y$.
Assume $0 \leq \ell(f(X), Y) \leq \ell_\infty$ for all $f \in \mathcal{F}$ and $(X, Y) \sim P$.

Then, with $\|f - f_i\| < \varepsilon$,

$$\begin{aligned}
\mathcal{R}[f] - \hat{\mathcal{R}}[f] &\leq \left|\mathcal{R}[f] - \mathcal{R}[f_i]\right| + \mathcal{R}[f_i] - \hat{\mathcal{R}}[f_i] + \left|\hat{\mathcal{R}}[f_i] - \hat{\mathcal{R}}[f]\right| \\
&\leq 2G\varepsilon + \max_{i=1,\dots,m(\varepsilon)} \left\{\mathcal{R}[f_i] - \hat{\mathcal{R}}[f_i]\right\}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left\{\mathcal{R}[f] - \hat{\mathcal{R}}[f]\right\}\right] &\leq 2G\varepsilon + \mathbb{E}\left[\max_{i=1,\dots,m(\varepsilon)} \left\{\mathcal{R}[f_i] - \hat{\mathcal{R}}[f_i]\right\}\right] \\
&\leq 2G\varepsilon + \sqrt{\frac{2\ell_\infty^2}{N} \log m(\varepsilon)}.
\end{aligned}$$

Example: Infinite models with covering number

For the sake of specificity[5], assume $m(\varepsilon) \sim \varepsilon^{-d}$. Choose $\varepsilon \sim 1/\sqrt{N}$.

Chaining things together, we get

$$\begin{aligned}\text{Estimation error} &= \mathcal{R}[\hat{f}] - \inf_{f' \in \mathcal{F}} \mathcal{R}[f'] \\ &\lesssim \frac{4G}{\sqrt{N}} + \sqrt{\frac{8\ell_\infty^2}{N}} \left(\sqrt{d \log(N)} + \sqrt{\log(2/\delta)} \right) + \text{Opt. error}\end{aligned}$$

with probability $1 - \delta$.

In many cases, the analysis is suboptimal. Rademacher complexity leads to sharper bounds.

[5]A compact set in $\mathbb{R}^d$ has $m(\varepsilon) \sim (\sqrt{d}/\varepsilon)^d$. Generally, when $\log m(\varepsilon) \sim d \log(\varepsilon)$ with logarithmic factors in d ignored, d is loosely considered to be the underlying "dimension" of $\mathcal{F}$.

Outline

Decision theory

Estimation error

Rademacher complexity

Example: Ball constrained linear prediction

Rademacher complexity

Let $\mathcal{H}$ be a class of $\mathbb{R}$-valued functions on $\mathcal{Z}$.
Let P be a probability distribution on $\mathcal{Z}$.

The *Rademacher complexity* of $\mathcal{H}$ is

$$\mathrm{Rad}_N(\mathcal{H}) = \mathop{\mathbb{E}}_{\substack{Z_1,\dots,Z_N \overset{\mathrm{iid}}{\sim} P \\ \varepsilon_1,\dots,\varepsilon_N \overset{\mathrm{iid}}{\sim} \mathrm{Rad}}} \left[\sup_{h\in\mathcal{H}} \frac{1}{N} \sum_{i=1}^N \varepsilon_i h(Z_i) \right],$$

where $\varepsilon_1,\dots,\varepsilon_N$ are Rademacher random variables, which are ± 1 w.p. $1/2$, and $Z_1,\dots,Z_N$ and $\varepsilon_1,\dots,\varepsilon_N$ are independent.

To clarify, $R_N(\mathcal{H})$ does depend on the distribution P, but we suppress the dependency on P for the sake of notational simplicity.

In general, $R_N(\mathcal{H})$ may not be well defined if $\sup_{h\in\mathcal{H}}$ leads to a non-measurable function. However, as far as I know, all practically parameterized function classes used in ML do not have this problem. (Countable supremum of measurable functions is measurable, and we can usually choose a countable dense subset of the parameters.)

Symmetrization technique

In the supervised learning setup, let $Z = (X, Y)$ and

$$h(Z) = \ell(f(X), Y), \qquad \mathcal{H} = \{\ell(f(x), y) \mid f \in \mathcal{F}\}.$$

So

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left\{ \mathcal{R}[f] - \hat{\mathcal{R}}[f] \right\} = \mathbb{E} \sup_{h \in \mathcal{H}} \left\{ \mathop{\mathbb{E}}_{Z \sim P} [h(Z)] - \frac{1}{N} \sum_{i=1}^{N} h(Z_i) \right\}.$$

Theorem

$$\mathbb{E} \sup_{h \in \mathcal{H}} \left\{ \mathop{\mathbb{E}}_{Z \sim P} [h(Z)] - \frac{1}{N} \sum_{i=1}^{N} h(Z_i) \right\} \leq 2\mathrm{Rad}_N(\mathcal{H})$$

and

$$\mathbb{E} \sup_{h \in \mathcal{H}} \left\{ \frac{1}{N} \sum_{i=1}^{N} h(Z_i) - \mathop{\mathbb{E}}_{Z \sim P} [h(Z)] \right\} \leq 2\mathrm{Rad}_N(\mathcal{H}).$$

Proof. We use the *symmetrization technique*, which introduces $Z'_1, \ldots, Z'_N \sim P$ as independent copies of $Z_1, \ldots, Z_N \sim P$ to write

$$\mathop{\mathbb{E}}_{Z \sim P} [h(Z)] = \mathop{\mathbb{E}}_{Z'_1, \ldots, Z'_N \sim P} \left[\frac{1}{N} \sum_{i=1}^{N} h(Z'_i) \right].$$

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{Z_1,\dots,Z_N\sim P}\left[\sup_{h\in\mathcal{H}}\left\{\mathop{\mathbb{E}}_{Z\sim P}[h(Z)]-\frac{1}{N}\sum_{i=1}^{N}h(Z_i)\right\}\right]\\
&=\mathop{\mathbb{E}}_{Z_1,\dots,Z_N\sim P}\left[\sup_{h\in\mathcal{H}}\left\{\mathop{\mathbb{E}}_{Z_1',\dots,Z_N'\sim P}\left[\frac{1}{N}\sum_{i=1}^{N}h(Z_i')\,\middle|\,Z_1,\dots,Z_N\right]-\frac{1}{N}\sum_{i=1}^{N}h(Z_i)\right\}\right]\\
&=\mathop{\mathbb{E}}_{Z_1,\dots,Z_N\sim P}\left[\sup_{h\in\mathcal{H}}\left\{\mathop{\mathbb{E}}_{Z_1',\dots,Z_N'\sim P}\left[\frac{1}{N}\sum_{i=1}^{N}h(Z_i')-\frac{1}{N}\sum_{i=1}^{N}h(Z_i)\,\middle|\,Z_1,\dots,Z_N\right]\right\}\right]\\
&\le\mathop{\mathbb{E}}_{Z_1,\dots,Z_N\sim P}\left[\mathop{\mathbb{E}}_{Z_1',\dots,Z_N'\sim P}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{N}\sum_{i=1}^{N}h(Z_i')-\frac{1}{N}\sum_{i=1}^{N}h(Z_i)\right\}\,\middle|\,Z_1,\dots,Z_N\right]\right]\\
&=\mathop{\mathbb{E}}_{\substack{Z_1,\dots,Z_N\sim P\\ Z_1',\dots,Z_N'\sim P}}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{N}\sum_{i=1}^{N}(h(Z_i')-h(Z_i))\right\}\right]\\
&\overset{(*)}{=}\mathop{\mathbb{E}}_{\substack{Z_1,\dots,Z_N\sim P\\ Z_1',\dots,Z_N'\sim P\\ \varepsilon_1,\dots,\varepsilon_N}}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i(h(Z_i')-h(Z_i))\right\}\right]\\
&\le\mathop{\mathbb{E}}_{\substack{Z_1,\dots,Z_N\sim P\\ Z_1',\dots,Z_N'\sim P\\ \varepsilon_1,\dots,\varepsilon_N}}\left[\sup_{h\in\mathcal{H}}\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i h(Z_i')+\sup_{h\in\mathcal{H}}\frac{1}{N}\sum_{i=1}^{N}(-\varepsilon_i)h(Z_i)\right]
\end{aligned}
$$

Symmetrization technique

$$
\begin{aligned}
&= \mathop{\mathbb{E}}_{\substack{Z'_1,\ldots,Z'_N\sim P\\ \varepsilon_1,\ldots,\varepsilon_N}}\left[\sup_{h\in\mathcal{H}}\frac{1}{N}\sum_{i=1}^N \varepsilon_i h(Z'_i)\right] + \mathop{\mathbb{E}}_{\substack{Z_1,\ldots,Z_N\sim P\\ \varepsilon_1,\ldots,\varepsilon_N}}\left[\sup_{h\in\mathcal{H}}\frac{1}{N}\sum_{i=1}^N \varepsilon_i h(Z_i)\right]\\
&= 2\mathop{\mathbb{E}}_{\substack{Z'_1,\ldots,Z'_N\sim P\\ \varepsilon_1,\ldots,\varepsilon_N}}\left[\sup_{h\in\mathcal{H}}\frac{1}{N}\sum_{i=1}^N \varepsilon_i h(Z'_i)\right]\\
&= 2\mathrm{Rad}_N(\mathcal{H}).
\end{aligned}
$$

The other bound

$$
\mathop{\mathbb{E}}_{Z_1,\ldots,Z_N\sim P}\left[\left\{\frac{1}{N}\sum_{i=1}^N h(Z_i) - \sup_{h\in\mathcal{H}}\mathop{\mathbb{E}}_{Z\sim P}[h(Z)]\right\}\right] \le 2\mathrm{Rad}_N(\mathcal{H}).
$$

follows from the same reasoning.

Symmetrization technique

We clarify the step

$$\mathop{\mathbb{E}}_{\substack{Z_1,\dots,Z_N\sim P\\ Z_1',\dots,Z_N'\sim P}}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{N}\sum_{i=1}^{N}(h(Z_i')-h(Z_i))\right\}\right]$$

$$\overset{(*)}{=} \mathop{\mathbb{E}}_{\substack{Z_1,\dots,Z_N\sim P\\ Z_1',\dots,Z_N'\sim P\\ \varepsilon_1,\dots,\varepsilon_N}}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i(h(Z_i')-h(Z_i))\right\}\right]$$

Since $Z_1,\dots,Z_N,Z_1',\dots,Z_N'$ are IID,

$$\begin{bmatrix} h(Z_1')-h(Z_1)\\ \vdots\\ h(Z_i')-h(Z_i)\\ \vdots\\ h(Z_N')-h(Z_N)\end{bmatrix} \overset{\mathcal{D}}{=} \begin{bmatrix} h(Z_1')-h(Z_1)\\ \vdots\\ h(Z_i)-h(Z_i')\\ \vdots\\ h(Z_N')-h(Z_N)\end{bmatrix}$$

for any $i=1,\dots,N$.

Symmetrization technique

For any (non-random) $\varepsilon_1, \ldots, \varepsilon_N \in \{-1, +1\}$, we have

$$\begin{bmatrix} h(Z_1') - h(Z_1) \\ \vdots \\ h(Z_i') - h(Z_i) \\ \vdots \\ h(Z_N') - h(Z_N) \end{bmatrix} \overset{\mathcal{D}}{=} \begin{bmatrix} \varepsilon_1(h(Z_1') - h(Z_1)) \\ \vdots \\ \varepsilon_i(h(Z_i) - h(Z_i')) \\ \vdots \\ \varepsilon_N(h(Z_N') - h(Z_N)) \end{bmatrix}$$

Therefore, for any (non-random) $\varepsilon_1, \ldots, \varepsilon_N \in \{-1, +1\}$, we have

$$\sup_{h \in \mathcal{H}} \left\{ \frac{1}{N} \sum_{i=1}^{N} (h(Z_i') - h(Z_i)) \right\} \overset{\mathcal{D}}{=} \sup_{h \in \mathcal{H}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i (h(Z_i') - h(Z_i)) \right\}$$

Taking the expectation with respect to Z, Z', and ε justifies $\overset{(*)}{=}$. $\square$

Contraction principle

Theorem

Let $a_1, \ldots, a_N$ and b be functions from Θ to $\mathbb{R}$ (no assumption). Let $\varphi_1, \ldots, \varphi_N$ be 1-Lipschitz functions from $\mathbb{R}$ to $\mathbb{R}$. Let $\varepsilon_1, \ldots, \varepsilon_N$ be IID Rademacher random variables. Then,

$$\mathop{\mathbb{E}}_{\varepsilon_1,\ldots,\varepsilon_N}\left[\sup_{\theta\in\Theta}\left\{b(\theta)+\sum_{i=1}^{N}\varepsilon_i\varphi_i(a_i(\theta))\right\}\right] \leq \mathop{\mathbb{E}}_{\varepsilon_1,\ldots,\varepsilon_N}\left[\sup_{\theta\in\Theta}\left\{b(\theta)+\sum_{i=1}^{N}\varepsilon_i a_i(\theta)\right\}\right].$$

Proof. Use induction. Statement holds trivially with $N = 0$.

Now assume statement holds for $N - 1$.

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\varepsilon_1,\ldots,\varepsilon_N}\Big[\sup_{\theta\in\Theta}\big\{b(\theta)+\sum_{i=1}^{N}\varepsilon_i\varphi_i(a_i(\theta))\big\}\Big]\\
&=\frac{1}{2}\mathop{\mathbb{E}}_{\varepsilon_1,\ldots,\varepsilon_{N-1}}\Big[\sup_{\theta\in\Theta}\big\{b(\theta)+\sum_{i=1}^{N-1}\varepsilon_i\varphi_i(a_i(\theta))+\varphi_N(a_N(\theta))\big\}\Big]\\
&\qquad+\frac{1}{2}\mathop{\mathbb{E}}_{\varepsilon_1,\ldots,\varepsilon_{N-1}}\Big[\sup_{\theta'\in\Theta}\big\{b(\theta')+\sum_{i=1}^{N-1}\varepsilon_i\varphi_i(a_i(\theta'))-\varphi_N(a_N(\theta'))\big\}\Big]\\
&=\mathop{\mathbb{E}}_{\varepsilon_1,\ldots,\varepsilon_{N-1}}\Big[\sup_{\theta,\theta'\in\Theta}\big\{\frac{b(\theta)+b(\theta')}{2}+\sum_{i=1}^{N-1}\varepsilon_i\frac{\varphi_i(a_i(\theta))+\varphi_i(a_i(\theta'))}{2}+\frac{\varphi_N(a_N(\theta))-\varphi_N(a_N(\theta'))}{2}\big\}\Big]\\
&\overset{(*)}{=}\mathop{\mathbb{E}}_{\varepsilon_1,\ldots,\varepsilon_{N-1}}\Big[\sup_{\theta,\theta'\in\Theta}\big\{\frac{b(\theta)+b(\theta')}{2}+\sum_{i=1}^{N-1}\varepsilon_i\frac{\varphi_i(a_i(\theta))+\varphi_i(a_i(\theta'))}{2}+\frac{|\varphi_N(a_N(\theta))-\varphi_N(a_N(\theta'))|}{2}\big\}\Big]\\
&\le\mathop{\mathbb{E}}_{\varepsilon_1,\ldots,\varepsilon_{N-1}}\Big[\sup_{\theta,\theta'\in\Theta}\big\{\frac{b(\theta)+b(\theta')}{2}+\sum_{i=1}^{N-1}\varepsilon_i\frac{\varphi_i(a_i(\theta))+\varphi_i(a_i(\theta'))}{2}+\frac{|a_N(\theta)-a_N(\theta')|}{2}\big\}\Big]\\
&\overset{(*)}{=}\mathop{\mathbb{E}}_{\varepsilon_1,\ldots,\varepsilon_{N-1}}\Big[\sup_{\theta,\theta'\in\Theta}\big\{\frac{b(\theta)+b(\theta')}{2}+\sum_{i=1}^{N-1}\varepsilon_i\frac{\varphi_i(a_i(\theta))+\varphi_i(a_i(\theta'))}{2}+\frac{a_N(\theta)-a_N(\theta')}{2}\big\}\Big]\\
&=\frac{1}{2}\mathop{\mathbb{E}}_{\varepsilon_1,\ldots,\varepsilon_{N-1}}\Big[\sup_{\theta\in\Theta}\big\{b(\theta)+\sum_{i=1}^{N-1}\varepsilon_i\varphi_i(a_i(\theta))+a_N(\theta)\big\}\Big]\\
&\qquad+\frac{1}{2}\mathop{\mathbb{E}}_{\varepsilon_1,\ldots,\varepsilon_{N-1}}\Big[\sup_{\theta'\in\Theta}\big\{b(\theta')+\sum_{i=1}^{N-1}\varepsilon_i\varphi_i(a_i(\theta'))-a_N(\theta')\big\}\Big]
\end{aligned}
$$

$\overset{(*)}{=}$ follows from considering the max over (θ, θ') and (θ', θ).

$$
\begin{aligned}
&= \frac{1}{2} \mathop{\mathbb{E}}_{\varepsilon_1,\dots,\varepsilon_{N-1}} \Big[\sup_{\theta\in\Theta} \Big\{ b(\theta) + \sum_{i=1}^{N-1} \varepsilon_i \varphi_i(a_i(\theta)) + a_N(\theta) \Big\} \Big] \\
&\quad + \frac{1}{2} \mathop{\mathbb{E}}_{\varepsilon_1,\dots,\varepsilon_{N-1}} \Big[\sup_{\theta'\in\Theta} \Big\{ b(\theta') + \sum_{i=1}^{N-1} \varepsilon_i \varphi_i(a_i(\theta')) - a_N(\theta') \Big\} \Big] \\
&= \mathop{\mathbb{E}}_{\varepsilon_N} \Bigg[\mathop{\mathbb{E}}_{\varepsilon_1,\dots,\varepsilon_{N-1}} \Big[\sup_{\theta\in\Theta} \Big\{ b(\theta) + \varepsilon_N a_N(\theta) + \sum_{i=1}^{N-1} \varepsilon_i \varphi_i(a_i(\theta)) \Big\} \Big] \,\Big|\, \varepsilon_N \Bigg] \\
&\leq \mathop{\mathbb{E}}_{\varepsilon_N} \Bigg[\mathop{\mathbb{E}}_{\varepsilon_1,\dots,\varepsilon_{N-1}} \Big[\sup_{\theta\in\Theta} \Big\{ b(\theta) + \varepsilon_N a_N(\theta) + \sum_{i=1}^{N-1} \varepsilon_i a_i(\theta) \Big\} \Big] \,\Big|\, \varepsilon_N \Bigg] \\
&= \mathop{\mathbb{E}}_{\varepsilon_1,\dots,\varepsilon_N} \Big[\sup_{\theta\in\Theta} \Big\{ b(\theta) + \sum_{i=1}^{N} \varepsilon_i a_i(\theta) \Big\} \Big],
\end{aligned}
$$

where the final inequality holds by the induction hypothesis. $\square$

Contraction principle: Corollary

Corollary

Let $\ell(\cdot, Y)$ be G-Lipschitz for all $Y \sim P_Y$. Let $\varepsilon_1, \ldots, \varepsilon_N$ be IID Rademacher random variables. Then,

$$\mathop{\mathbb{E}}_{\varepsilon_1,\ldots,\varepsilon_N}\left[\sup_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} \varepsilon_i \ell(f(X_i), Y_i) \,\middle|\, \{(X_i, Y_i)\}_{i=1}^N\right]$$
$$\leq G \cdot \mathop{\mathbb{E}}_{\varepsilon_1,\ldots,\varepsilon_N}\left[\sup_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} \varepsilon_i f(X_i) \,\middle|\, \{(X_i, Y_i)\}_{i=1}^N\right].$$

Taking expectation with respect to $\{(X_i, Y_i)\}_{i=1}^N$, we conclude

$$\mathrm{Rad}_N(\mathcal{H}) \leq G \cdot \mathrm{Rad}_N(\mathcal{F}).$$

To be pedantic, we should write

$$\mathrm{Rad}_N(\mathcal{H}; P_{X,Y}) \leq G \cdot \mathrm{Rad}_N(\mathcal{F}; P_X),$$

Since the LHS depends on the joint distribution $P_{X,Y}$ while the RHS depends only on the marginal distribution P_X.

Outline

Decision theory

Estimation error

Rademacher complexity

Example: Ball constrained linear prediction

Ball constrained linear prediction

Let

$$\mathcal{F} = \left\{ f_\theta(x) = \theta^\intercal x \,\middle|\, \|\theta\| \le D,\ \theta \in \mathbb{R}^d \right\},$$

where $\|\cdot\|$ is some norm. Then,

$$\begin{aligned}
\mathrm{Rad}_N(\mathcal{F}) &= \mathop{\mathbb{E}}_{\substack{X_1,\ldots,X_N \overset{\text{iid}}{\sim} P_X \\ \varepsilon_1,\ldots,\varepsilon_N \overset{\text{iid}}{\sim} \mathrm{Rad}}} \left[\sup_{\|\theta\| \le D} \frac{1}{N} \sum_{i=1}^N \varepsilon_i \theta^\intercal X_i \right] = \mathop{\mathbb{E}}_{\substack{X_1,\ldots,X_N \\ \varepsilon_1,\ldots,\varepsilon_N}} \left[\sup_{\|\theta\| \le D} \frac{1}{N} \varepsilon^\intercal \mathbf{X} \theta \right] \\
&= \frac{D}{N} \mathop{\mathbb{E}}_{\substack{X_1,\ldots,X_N \\ \varepsilon_1,\ldots,\varepsilon_N}} \left[\sup_{\|\theta\| \le 1} \theta^\intercal (\mathbf{X}^\intercal \varepsilon) \right] = \frac{D}{N} \mathop{\mathbb{E}}_{\substack{X_1,\ldots,X_N \\ \varepsilon_1,\ldots,\varepsilon_N}} [\|\mathbf{X}^\intercal \varepsilon\|_*],
\end{aligned}$$

where $\|\cdot\|_*$ denotes the dual norm and

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{bmatrix} \in \mathbb{R}^N, \qquad \mathbf{X} = \begin{bmatrix} X_1^\intercal \\ \vdots \\ X_N^\intercal \end{bmatrix} \in \mathbb{R}^{N \times d}.$$

Euclidean norm case

Assume $\|X\|_2 \leq R$ for all $X \sim P_X$. When $\|\cdot\| = \|\cdot\|_* = \|\cdot\|_2$,

$$\begin{aligned}
\mathrm{Rad}_N(\mathcal{F}) &= \frac{D}{N}\mathbb{E}[\|\mathbf{X}^\intercal \varepsilon\|_2] \leq \frac{D}{N}\sqrt{\mathbb{E}[\|\mathbf{X}^\intercal \varepsilon\|_2^2]} \\
&= \frac{D}{N}\sqrt{\mathbb{E}[\mathrm{Tr}(\varepsilon^\intercal \mathbf{X}\mathbf{X}^\intercal \varepsilon)]} = \frac{D}{N}\sqrt{\mathbb{E}[\mathrm{Tr}(\mathbf{X}\mathbf{X}^\intercal \varepsilon\varepsilon^\intercal)]} = \frac{D}{N}\sqrt{\mathbb{E}[\mathrm{Tr}(\mathbf{X}\mathbf{X}^\intercal I)]} \\
&= \frac{D}{N}\sqrt{\sum_{i=1}^{N}\mathbb{E}[\|X_i\|_2^2]} = \frac{D}{\sqrt{N}}\sqrt{\mathop{\mathbb{E}}_{X\sim P}[\|X\|_2^2]} \\
&\leq \frac{DR}{\sqrt{N}},
\end{aligned}$$

where we used Jensen's inequality and the trace trick.

ℓ_1-ℓ_∞-norm case

Assume $\|X\|_\infty \leq R$ for all $X \sim P_X$. When $\|\cdot\| = \|\cdot\|_1$ and $\|\cdot\|_* = \|\cdot\|_\infty$,

$$\begin{aligned}\mathrm{Rad}_N(\mathcal{F}) &= \frac{D}{N}\mathbb{E}[\|\mathbf{X}^\intercal \varepsilon\|_\infty] \\ &= \frac{D}{N}\mathbb{E}\Big[\max_{j=1,\ldots,d}\Big|\sum_{i=1}^{N}(X_i)_j\varepsilon_i\Big|\Big] \\ &\leq \frac{DR}{\sqrt{N}}\sqrt{2\log(2d)},\end{aligned}$$

since $(X_i)_j\varepsilon_i \in [-R, R]$ is a sub-Gaussian with $\tau = R$, and the sum of N such sub-Gaussians is a sub-Gaussian with $\tau = \sqrt{N}R$.

Estimation error

Let $\|\cdot\|$ be the Euclidean norm. Assume $\|X\| \le R$ for all $X \sim P_X$. Assume $\ell(\cdot, Y)$ is G-Lipschitz for all $Y \sim P_Y$. Then,

$$\begin{aligned}
\mathbb{E}[\mathcal{R}[f_{\hat{\theta}}]] - \inf_{\|\theta\| \le D} \mathcal{R}[f_\theta] &\le \mathbb{E} \sup_{f \in \mathcal{F}} \{\mathcal{R}[f] - \hat{\mathcal{R}}[f]\} + \mathbb{E} \sup_{f \in \mathcal{F}} \{\hat{\mathcal{R}}[f] - \mathcal{R}[f]\} \\
&\quad + \underbrace{\mathbb{E}\,(\hat{\mathcal{R}}[\hat{f}] - \inf_{f \in \mathcal{F}} \hat{\mathcal{R}}[f])}_{=\text{Opt. error}} \\
&\le 4\mathrm{Rad}_N(\mathcal{H}) + \text{Opt. error} \\
&\le 4G\mathrm{Rad}_N(\mathcal{F}) + \text{Opt. error} \\
&\le \frac{4DGR}{\sqrt{N}} + \text{Opt. error.}
\end{aligned}$$

The first ineq. is by the estimation error decomposition, the second by the symmetrization technique, and the third by the contraction principle.