Machine Learning Theory (CSC 482A/581B) - Lectures 5 and 6

Nishant Mehta

1 Agnostic Learning

From now on, we will study the agnostic learning setting, wherein the labels themselves can be random conditional on the input. Thus, the distribution P is now a joint distribution over $\mathcal{X} \times \mathcal{Y}$. In PAC learning, it made sense to analyze the rate of convergence of a learning algorithm's risk $R(\hat{f})$ to zero as the sample size n increases; however, in the agnostic setting it may no longer be the case that there exists a hypothesis from the set $\bar{\mathcal{F}}$ of all possible hypotheses that obtains zero risk.[1] A more sensible goal is to hope for a learning algorithm for which the *excess risk* with respect to the best possible hypothesis decays to zero as $n \to \infty$. Let's therefore study the behavior of this best possible hypothesis.

1.1 Bayes classifier

Definition 1. The Bayes risk R^* is defined as the minimum risk among all possible hypotheses:[2]

$$R^* = \inf_{f \in \bar{\mathcal{F}}} R(f).$$

A *Bayes optimal classifier*, or Bayes classifier, is a hypothesis f which obtains the Bayes risk:

$$R(f) = R^*.$$

What form does a Bayes classifier take? For an input $x \in \mathcal{X}$, it is easy to see that the conditional risk $\mathsf{E}[\mathbf{1}[\hat{y} \neq Y] \mid X = x] = \Pr(\hat{y} \neq Y \mid X = x)$ is minimized by predicting

$$\hat{y} \in \operatorname*{arg\,max}_{y \in \{0,1\}} \Pr(Y = y \mid X = x).$$

Hence, by arbitrarily breaking ties in favor of the positive class, we take the Bayes classifier to be

$$f_{\text{Bayes}}(x) = \mathbf{1}[\Pr(Y = 1 \mid X = x) \geq 1/2] \, .$$

1.2 Minimizing excess risk with respect to f_{Bayes} is hopeless

Now that we have defined the excess risk with respect to f_{Bayes}, a natural question arises:

> For a fixed input space $\mathcal{X}$, is there a learning algorithm $\mathcal{A}$ for which, no matter the distribution P, for any $\varepsilon > 0$ and $\delta \in (0, 1)$, there is a sample size $n(\varepsilon, \delta)$ (not depending on P) such that $R(\hat{f}) - R^* \leq \varepsilon$ with probability at least $1 - \delta$?

[1] Technically, whenever I refer to the set of all possible hypotheses, I actually mean the set of all measurable hypotheses. If you do not know what the term "measurable" means, do not worry about this footnote.

[2] If you do not know what inf means: if you are an undergrad, inf stands for "infimum", and you may think of it roughly as "minimum". If you are a graduate student, you should familiarize yourself with infimums and supremums.

Such a learning algorithm would be a universal learner, as it performs well against *any* distribution P. Unfortunately, this is not possible, as shown by the following "No-Free-Lunch" result:

> In agnostic learning, for any learning algorithm and sample size n, there is a distribution $P \in \Delta(\mathcal{X} \times \mathcal{Y})$ with deterministic labels under which, with constant positive probability, $R(\hat{f})$ is lower bounded by a positive constant.

Remarks:

- Since the label Y is deterministic given X, we have $P(Y = 1 \mid X) \in \{0, 1\}$; consequently, there is a perfect labeling function, and hence $R(\hat{f}) - R^* = R(\hat{f})$.

1.3 The Agnostic Model

In light of the impossibility result of competing with hypothesis f_{Bayes}, we will instead compete against the best hypothesis within our hypothesis space $\mathcal{F}$.

Let f^* be a hypothesis in $\mathcal{F}$ that minimizes the risk (under distribution P), so that

$$R(f^*) = \inf_{f \in \mathcal{F}} R(f).$$

Definition 2. We say that $\mathcal{F}$ is *agnostically learnable* if there exists an algorithm $\mathcal{A}$ and a function $n \colon (0,1)^2 \to \mathbb{N}$ which, for any distribution P over $\mathcal{X} \times \mathcal{Y}$ and for all $\varepsilon \in (0,1)$ and $\delta \in (0,1)$, satisfy the following guarantee:

If $\mathcal{A}$ is given access to $n(\varepsilon, \delta)$ labeled examples drawn i.i.d. from P, then with probability at least $1 - \delta$, $\mathcal{A}$ outputs a hypothesis $\hat{f}$ with excess risk $R(f) - R(f^*) \leq \varepsilon$.

We say that $\mathcal{F}$ is *efficiently agnostically learnable* if, in addition, $\mathcal{A}$ runs in time polynomial in $\frac{1}{\varepsilon}$ and $\frac{1}{\delta}$.

As with PAC learning, we can also require $\mathcal{A}$ to output hypotheses in $\mathcal{F}$ (so that $\mathcal{A}$ is proper). If $\mathcal{F}$ is agnostically learnable by such an algorithm, then $\mathcal{F}$ is *proper agnostically learnable*.

1.4 Error decompositions

Decomposition of excess risk into approximation error and estimation error

Consider a learning algorithm $\mathcal{A}$ which, given a training sample S, outputs some hypothesis $\hat{f} \in \mathcal{F}$. The excess risk of $\hat{f}$ with respect to the Bayes classifier can be decomposed as

$$R(\hat{f}) - R^* = \underbrace{\left(R(f^*) - R^*\right)}_{\text{approximation error}} + \underbrace{\left(R(\hat{f}) - R(f^*)\right)}_{\text{estimation error}}. \tag{1}$$

The first term in the decomposition is the *approximation error*: it is a measure of how well the class $\mathcal{F}$ can approximate the Bayes classifier in terms of risk; if $f_{\text{Bayes}} \in \mathcal{F}$, then the approximation error is zero. Bounding the approximation error requires knowledge of R^* or some partial information about f_{Bayes}, and in settings where little or no distributional assumptions are made, it is thus very difficult if not impossible to control the approximation error. Note, however, that if we begin making certain assumptions about P and if we allow $\mathcal{F}$ to grow with n (so that at sample size n Learner outputs a hypothesis in $\mathcal{F}_n$), then it is possible to obtain rates of convergence of the approximation error to zero.

The second term in the decomposition is the *estimation error*. Unlike the approximation error, provided that $\mathcal{F}$ is not "too large" it is possible to obtain good bounds on the estimation error in a distribution-free way, i.e. without having any information about the underlying distribution P. How does the estimation error typically depend on $\mathcal{F}$? As we will soon see, for learning algorithms that return hypotheses that have low empirical risk, the estimation error increases with $|\mathcal{F}|$. This accords with our intuition that, information-theoretically, we need more bits of information to "whittle down" $\mathcal{F}$ to the risk minimizer (or set of risk minimizers) as $\mathcal{F}$ increases in size.

The decomposition (1) into approximation error and estimation error highlights the familiar trade-off between *model expressivity* and *generalization*. As we increase the size (complexity) of our model $\mathcal{F}$, the approximation error decreases since the model can express more patterns; simultaneously, however, it becomes more likely that we will overfit and hence fail to generalize well.

Our primary focus will be controlling the estimation error. Controlling the estimation error rather than the excess risk with respect to f_{Bayes} has various motivations, including

- If we are "lucky" and the approximation error is zero or sufficiently small, a bound on the estimation error also provides a good bound on $R(\hat{f}) - R^*$.
- Suppose that we are in a nonparametric setup where, at sample size n, Learner employs hypothesis space $\mathcal{F}_n$. Under mild assumptions about the true distribution, we may be able to control the approximation error as a function of n. It then is also useful to control the estimation error for each $\mathcal{F}_n$, as we then can determine how quickly the complexity of the model should increase with the sample size.

Oracle inequality approach. A bound on the estimation error of a learning algorithm $\mathcal{A}$ that outputs hypothesis $\hat{f}$ is equivalent to a bound of the form:

$$R(\hat{f}) \leq R(f^*) + \textsc{bound}(\mathcal{F}, n). \tag{2}$$

In statistics and machine learning, a bound of this form is called an *oracle inequality*. The name stems from our comparing the performance of $\hat{f}$ to that of an omniscient oracle which plays f^*, the best hypothesis in $\mathcal{F}$.

It is natural to seek an oracle inequality for a learning algorithm, as we then know how far off the risk we obtain is from the best possible risk obtainable via $\mathcal{F}$. However, to the practitioner, oracle inequalities are not immediately useful: $R(f^*)$ is an unknown quantity, so, while the bound may be correct, a practitioner has no observable upper bound on $R(\hat{f})$ (!).

Deviations approach: Decomposition of risk into empirical risk and deviation

The error decomposition below, this time of the risk of $R(\hat{f})$ itself, *can* lead to an observable bound. For any hypothesis f and training sample $S = ((X_1, Y_1), \ldots, (X_n, Y_n))$, let $\hat{R}_S(f) = \frac{1}{n}\sum_{j=1}^n \mathbf{1}[f(X_j) \neq Y_j]$ denote the empirical risk of f on S. Then

$$R(\hat{f}) = \hat{R}_S(\hat{f}) + \underbrace{\left(R(\hat{f}) - \hat{R}_S(\hat{f})\right)}_{\text{deviation}}. \tag{3}$$

Let's see how we can use (3) to get an upper bound on the risk of $\hat{f}$. Suppose that we have a bound of the form:

$$\left| R(f) - \hat{R}_S(f) \right| \leq \varepsilon \qquad \text{for all } f \in \mathcal{F}. \tag{4}$$

This bound is known as a uniform deviation bound, since it bounds the deviation of $\hat{R}_S(f)$ from its mean $\mathsf{E}[\hat{R}_S(f)] = R(f)$, uniformly over $\mathcal{F}$.

From (4), the bound holds for $\hat{f}$ in particular, and so we immediately obtain the risk bound

$$R(\hat{f}) \leq \hat{R}_S(\hat{f}) + \varepsilon. \tag{5}$$

Note that this upper bound is observable, since the empirical risk of $\hat{f}$ can be observed.

Moreover, as we will see below, a bound of the form (4) can, with just a few short steps, lead to an oracle inequality.

2 A first excess risk bound for finite classes

Let's derive a first excess risk bound for agnostically learning a finite class $\mathcal{F}$. We will obtain a bound by way of a concentration inequality known as Hoeffding's inequality, proved by Wassily Hoeffding in 1963.

Theorem 1. *Let $Z_1, \ldots, Z_n$ be independent random variables such that $Z_j \in [a_j, b_j]$ for $j \in [n]$. Let $\bar{Z} = \frac{1}{n} \sum_{j=1}^n Z_j$. Then for any $\varepsilon > 0$:*

$$\Pr\left(\bar{Z} - \mathsf{E}[\bar{Z}] \geq \varepsilon\right) \leq \exp\left(\frac{-2n^2\varepsilon^2}{\sum_{j=1}^n (b_j - a_j)^2}\right).$$

Before establishing an excess risk bound, we will first establish a uniform convergence result: the empirical risk converges to the actual risk uniformly over $\mathcal{F}$. It becomes tiresome to carry around the subscript S for the empirical risk, so we use the abbreviation $\hat{R}(f) := \hat{R}_S(f)$.

Theorem 2. *Let $\mathcal{F}$ be a finite set of hypotheses and let P be a fixed distribution over $\mathcal{X} \times \mathcal{Y}$. For any $\varepsilon > 0$ and any $\delta \in (0, 1)$, if $(X_1, Y_1), \ldots, (X_n, Y_n)$ are drawn i.i.d. from P with*

$$n \geq \frac{\log |\mathcal{F}| + \log \frac{2}{\delta}}{2\varepsilon^2},$$

then with probability at least $1 - \delta$

$$\left| R(f) - \hat{R}(f) \right| \leq \varepsilon \quad \textit{for all } f \in \mathcal{F}.$$

Proof. Fix some $f \in \mathcal{F}$ and consider the probability that

$$R(f) - \hat{R}(f) > \varepsilon.$$

This event may be rewritten as

$$\frac{1}{n} \sum_{j=1}^n \mathbf{1}[f(X_j) \neq Y_j] - \mathsf{E}\left[\frac{1}{n} \sum_{j=1}^n \mathbf{1}[f(X_j) \neq Y_j]\right] > \varepsilon,$$

and so we may apply Hoeffding's inequality twice, once with $Z_j = -\mathbf{1}[f(X_j) \neq Y_j]$, $a_j = 0$, and $b_j = 1$ for $j \in [n]$, yielding

$$\Pr\left(R(f) - \hat{R}(f) > \varepsilon\right) \leq e^{-2n\varepsilon^2},$$

and once with $Z_j = \mathbf{1}[f(X_j) \neq Y_j]$, $a_j = -1$, and $b_j = 0$ for $j \in [n]$, yielding

$$\Pr\left(\hat{R}(f) - R(f) > \varepsilon\right) \leq e^{-2n\varepsilon^2}.$$

Hence,

$$\Pr\left(\left|R(f) - \hat{R}(f)\right| > \varepsilon\right) \leq 2e^{-2n\varepsilon^2}.$$

Next, applying the union bound, we have

$$\begin{aligned}\Pr\left(\exists f \in \mathcal{F} : \left|R(f) - \hat{R}(f)\right| > \varepsilon\right) &\leq \sum_{f\in\mathcal{F}} \Pr\left(\left|R(f) - \hat{R}(f)\right| > \varepsilon\right) \\ &\leq 2|\mathcal{F}|e^{-2n\varepsilon^2}.\end{aligned}$$

The result follows by setting the RHS to δ and solving for n. □

We now prove that any finite class can be agnostically learned using *empirical risk minimization* (ERM) over $\mathcal{F}$, a method which outputs the hypothesis in $\mathcal{F}$ that minimizes the empirical risk.

Theorem 3. *Let $\mathcal{F}$ be a finite set of hypotheses, let P be a fixed distribution over $\mathcal{X} \times \mathcal{Y}$, and take $\mathcal{A}$ to be ERM over $\mathcal{F}$ For any $\varepsilon > 0$ and any $\delta \in (0, 1)$, if $\mathcal{A}$ is run on a training sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ drawn i.i.d. from P with*

$$n \geq \frac{2\left(\log|\mathcal{F}| + \log\frac{2}{\delta}\right)}{\varepsilon^2},$$

then with probability at least $1 - \delta$

$$R(\hat{f}) \leq R(f^*) + \varepsilon.$$

Proof. First, observe that

$$\begin{aligned}R(\hat{f}) - R(f^*) = &\ \left(\hat{R}(\hat{f}) + (R(\hat{f}) - \hat{R}(\hat{f}))\right) \\ &- \left(\hat{R}(f^*) + (R(f^*) - \hat{R}(f^*))\right) \\ = &\ \left(\hat{R}(\hat{f}) - \hat{R}(f^*)\right) + \left(R(\hat{f}) - \hat{R}(\hat{f})\right) + \left(\hat{R}(f^*) - R(f^*)\right) \\ \leq &\ \left(R(\hat{f}) - \hat{R}(\hat{f})\right) + \left(\hat{R}(f^*) - R(f^*)\right) \\ \leq &\ 2\max_{f\in\mathcal{F}}\left|R(f) - \hat{R}(f)\right|,\end{aligned}$$

where the first inequality uses the fact that the empirical risk of ERM is no greater than the empirical risk of f^*. Next, from Theorem 2, with probability at least $1 - \delta$

$$\max_{f\in\mathcal{F}}\left|R(\hat{f}) - \hat{R}(\hat{f})\right| \leq \varepsilon/2,$$

and so the result holds. □

3 Effective size of a class

So far, we have seen how to obtain a uniform convergence result when $\mathcal{F}$ is finite. We will now "upgrade" this result to the case when $\mathcal{F}$ is infinite. It is worth thinking about whether our previous proof might already yield a useful bound for infinite $\mathcal{F}$. Unfortunately, the answer is no because the union bound for infinite $\mathcal{F}$ leads to an infinite upper bound. As it turns out, the right way to derive a good uniform convergence bound still relies on a union bound, but applied in a very clever way. For this, we need the notion of the "effective size" of $\mathcal{F}$.

A key idea we will use is that even though $\mathcal{F}$ may be infinite, there are only finitely many ways to classify a given training sample by picking different hypotheses from $\mathcal{F}$. Let's make this concrete. Given a sequence of inputs $\mathbf{x}_1^n = (x_1, \ldots, x_n)$, let $\mathcal{F}_{|\mathbf{x}_1^n}$ be the coordinate projection of $\mathcal{F}$ onto $\mathbf{x}_1^n$. That is,

$$\mathcal{F}_{\mathbf{x}_1^n} := \left\{ \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{pmatrix} : f \in \mathcal{F} \right\}.$$

Since $\mathcal{F}$ is a set of classifiers, each of which takes values in $\{0, 1\}$, we have that $\mathcal{F}_{\mathbf{x}_1^n} \subset \{0, 1\}^n$, and hence $|\mathcal{F}_{\mathbf{x}_1^n}| \leq 2^n$.

Intuitively, even though our hypothesis space $\mathcal{F}$ is infinite, when it is viewed through the lens of the data, there are only finitely many distinct hypotheses.

Example 1 (Threshold functions)**.** Consider learning threshold functions over $\mathbb{R}$, so that $\mathcal{X} = \mathbb{R}$ and $\mathcal{F} = \{f_t : t \in \mathbb{R}\}$, where $f_t(x) = \mathbf{1}[x \geq t]$. Suppose that we have n distinct inputs $x_1 < x_2 < \ldots < x_n$. Then it is easy to see that there only $n + 1$ distinct ways that $\mathcal{F}$ can classify this training sample, namely:

$$t \in (-\infty, x_1) \quad t \in (x_1, x_2) \quad t \in (x_2, x_3) \quad \cdots \quad t \in (x_{n-1}, x_n) \quad t \in (x_n, \infty).$$

Thus, in this case, $|\mathcal{F}_{|\mathbf{x}_1^n}| = n + 1$, and for *any* training sample of size n, $|\mathcal{F}_{|\mathbf{x}_1^n}| \leq n + 1$.

4 Growth function

Definition 3. The *growth function* of $\mathcal{F}$ is defined as

$$\Pi_{\mathcal{F}}(n) = \sup_{(x_1, \ldots, x_n) \in \mathcal{X}^n} |\mathcal{F}_{|\mathbf{x}_1^n}|.$$

The growth function of $\mathcal{F}$ evaluated at n is the maximum number of ways a set of n points can be split using functions from $\mathcal{F}$. Often in the literature, you may also see $\Pi_{\mathcal{F}}(n)$ called the n^{th} shatter coefficient of $\mathcal{F}$. The latter term stems from the notion of "shattering", a crucial component in defining the fundamental notion of Vapnik-Chervonenkis dimension (VC dimension); we will study both shattering and VC dimension next week.

Example 2 (Intervals)**.** Consider the class of intervals over $\mathbb{R}$. We thus have $\mathcal{X} = \mathbb{R}$ and $\mathcal{F} = \{f_{a,b} : -\infty \leq a \leq b \leq \infty\}$ for $f_{a,b}(x) = \mathbf{1}[a \leq x \leq b]$. Similar to the case of threshold functions over $\mathbb{R}$, in this case $\Pi_{\mathcal{F}}(n)$ is finite. We leave it as exercise to work out the exact value of $\Pi_{\mathcal{F}}(n)$.

5 Uniform convergence for infinite classes

We now prove the following fundamental result, originally proved by Vapnik and Chervonenkis in 1971. The result is a uniform convergence result over infinite classes, where the complexity of the class is paid for via the growth function $\Pi_{\mathcal{F}}(n)$.

Before presenting the result, we establish some useful notation that appears frequently in the empirical process theory literature. For any function g from $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$, let $P\,g$ denote the expectation of $g(Z)$ for $Z = (X, Y) \sim P$. For a training sample $Z_1, \ldots, Z_n$, the empirical distribution P_n (with respect to $Z_1, \ldots, Z_n$) is $\frac{1}{n} \sum_{j=1}^n \delta_{Z_j}(\cdot)$. Let $P_n\,g$ denote the expectation of $g(Z)$ when Z is drawn from the empirical distribution P_n. We thus have

$$P\,g = \mathsf{E}_{Z \sim P}[g(Z)] \qquad\qquad P_n\,g = \frac{1}{n} \sum_{j=1}^n g(Z_j).$$

For any hypothesis $f \in \mathcal{F}$, define the loss-composed version of f as $g_f \colon (x, y) \mapsto \mathbf{1}[f(x) \neq y]$. So, while f is a random variable mapping from $\mathcal{X}$ to $\{0,1\}$, the function g_f is a random variable mapping from $\mathcal{X} \times \mathcal{Y}$ to $\{0,1\}$. From $\mathcal{F}$ we can generate the corresponding "loss-composed" class $\mathcal{G} = \{g_f : f \in \mathcal{F}\}$. Throughout this section, we will use the notation $Z = (X, Y)$ and $Z_j = (X_j, Y_j)$.

Theorem 4 (Vapnik and Chervonenkis, 1971). *For any probability distribution P and any hypothesis space $\mathcal{F} \subset \{0,1\}^{\mathcal{X}}$, and any $\varepsilon > 0$,*

$$\Pr\left(\sup_{g \in \mathcal{G}} |(P - P_n)g| > \varepsilon\right) \leq 8\Pi_{\mathcal{F}}(n) e^{-n\varepsilon^2/32}.$$

The proof of this result uses two key ideas, both of which are variants of a powerful argument known as symmetrization.

The first symmetrization is sometimes called "symmetrization by ghost sample". The idea is to shift from bounding the uniform deviation of an empirical expectation from the actual expectation to bounding the uniform deviation between two empirical expectations from independent samples of the same size. To this end, let $Z'_1, \ldots, Z'_n$ be an independent copy of $Z_1, \ldots, Z_n$, so that all $2n$ random variables are i.i.d. according to P. We call $Z'_1, \ldots, Z'_n$ a ghost sample, because this is a fictional sample that we do not actually have, but which, nevertheless, we will use in our analysis. Now, similar to P_n, let $P'_n\,g$ denote the empirical expectation of a function $g \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ with respect to the ghost sample, so that

$$P'_n\,g = \frac{1}{n} \sum_{j=1}^n g(Z'_j).$$

With this notation in place, we establish the first symmetrization lemma.

Lemma 1. *Let $Z_1, \ldots, Z_n, Z'_1, \ldots, Z'_n$ be i.i.d. random variables distributed according to P. Then for any ε satisfying $n\varepsilon^2 \geq 2$,*

$$\Pr\left(\sup_{g \in \mathcal{G}} |(P - P_n)g| > \varepsilon\right) \leq 2 \Pr\left(\sup_{g \in \mathcal{G}} |(P'_n - P_n)g| > \varepsilon/2\right)$$

Intuitively, the reason why we view this lemma as progress is because we now only care about viewing the function class $\mathcal{G}$ through the lens of a double sample (the original sample and the ghost sample). Thus, it is conceivable that we may be able to use the finiteness of $\Pi_{\mathcal{F}}(2n)$ if we are clever enough.

Proof. The proof involves a sequence of lower bounds on the probability in the RHS of the lemma.

Let g_n be a function in $\mathcal{G}$ for which $|(P - P_n)g_n| = \sup_{g\in\mathcal{G}} |(P - P_n)g|$ (if there is no such g_n, minor but tedious modifications allow essentially the same proof to go through). Then

$$\Pr\left(\sup_{g\in\mathcal{G}} |(P_n - P_n')g| > \varepsilon/2\right) \geq \Pr(|(P_n - P_n')g_n| > \varepsilon/2). \tag{6}$$

Next, observe that $[|(P - P_n)g_n| > \varepsilon]$ and $[|(P - P_n')g_n| < \frac{\varepsilon}{2}]$ together imply $[|(P_n' - P_n)g_n| > \frac{\varepsilon}{2}]$. Hence, the above is at least

$$\begin{aligned}&\Pr\Big((|(P - P_n)g_n| > \varepsilon) \bigwedge (|(P - P_n')g_n| < \varepsilon/2)\Big)\\ &= \mathsf{E}\Big[\mathbf{1}[|(P - P_n)g_n| > \varepsilon] \cdot \Pr\left(|(P - P_n')g_n| < \varepsilon/2 \mid Z_1, \ldots, Z_n\right)\Big]\end{aligned} \tag{7}$$

Now, since g_n depends only on $Z_1, \ldots, Z_n$, we can lower bound the conditional probability above using Chebyshev's inequality:

$$\begin{aligned}\Pr_{Z_1',\ldots,Z_n'}\left((P - P_n')g_n \geq \varepsilon/2\right) &\leq \frac{\mathrm{Var}[g_n(Z_1') \mid Z_1, \ldots, Z_n]}{n\varepsilon^2/4}\\ &\leq \frac{1}{n\varepsilon^2},\end{aligned}$$

where we used the fact that the variance of a Bernoulli random variable can be at most 1/4. Since we assumed that $n\varepsilon^2 \geq 2$, it follows that

$$\Pr\left(|(P - P_n')g_n| < \varepsilon/2 \mid Z_1, \ldots, Z_n\right) \geq \frac{1}{2}.$$

But this implies that (7) (and hence the LHS of (6)) is lower bounded by

$$\frac{1}{2}\Pr\left(|(P - P_n)g_n| > \varepsilon\right).$$

□

The second key idea is another application of symmetrization, this time an argument often called "symmetrization by random signs". For this argument, we employ a sequence of independent Rademacher random variables $\sigma_1, \ldots, \sigma_n$. A Rademacher random variable σ is one which takes the values $\{-1, +1\}$ with equal probability, so $\Pr(\sigma = -1) = \Pr(\sigma = 1) = \frac{1}{2}$.

Lemma 2. *Let $Z_1, \ldots, Z_n, Z_1', \ldots, Z_n'$ be i.i.d. random variables and let $\sigma_1, \ldots, \sigma_n$ be independent Rademacher random variables. Then for any $\varepsilon > 0$,*

$$\Pr\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{j=1}^{n}\Big(g(Z_j') - g(Z_j)\Big)\right| > \varepsilon/2\right) \leq 2\Pr\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{j=1}^{n}\sigma_j g(Z_j)\right| > \varepsilon/4\right)$$

Proof. Observe that for any choice of sign variables $\sigma_1, \ldots, \sigma_n \in \{-1, +1\}$, the distribution of

$$\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{j=1}^{n}\Big(g(Z_j') - g(Z_j)\Big)\right|$$

is identical to the distribution of

$$\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{j=1}^{n}\sigma_j\left(g(Z_j')-g(Z_j)\right)\right|.$$

Therefore, letting $\sigma_1,\ldots,\sigma_n$ be i.i.d. Rademacher random variables, it holds that

$$\begin{aligned}
&\Pr\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{j=1}^{n}\left(g(Z_j')-g(Z_j)\right)\right|>\varepsilon/2\right)\\
&=\Pr\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{j=1}^{n}\sigma_j\left(g(Z_j')-g(Z_j)\right)\right|>\varepsilon/2\right)\\
&\le\Pr\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{j=1}^{n}\sigma_j g(Z_j')\right|>\varepsilon/4\right)+\Pr\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{j=1}^{n}\sigma_j g(Z_j)\right|>\varepsilon/4\right)\\
&=2\Pr\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{j=1}^{n}\sigma_j g(Z_j)\right|>\varepsilon/4\right)
\end{aligned}$$

□

Proof (of Theorem 4). Lemmas 1 and 2 together imply that for $n\varepsilon^2\ge 2$,

$$\Pr\left(\sup_{g\in\mathcal{G}}|(P-P_n)g|>\varepsilon\right)\le 4\Pr\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{j=1}^{n}\sigma_j g(Z_j)\right|>\varepsilon/4\right).$$

Next, observe that

$$\begin{aligned}
&\Pr\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{j=1}^{n}\sigma_j g(Z_j)\right|>\varepsilon/4\mid Z_1,\ldots,Z_n\right)\\
&=\Pr\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{j=1}^{n}\sigma_j g_f(Z_j)\right|>\varepsilon/4\mid Z_1,\ldots,Z_n\right)\\
&=\Pr\left(\max_{v\in\mathcal{F}_{|\mathbf{x}_1^n}}\left|\frac{1}{n}\sum_{j=1}^{n}\sigma_j\,\mathbf{1}[v_j\ne Y_j]\right|>\varepsilon/4\mid Z_1,\ldots,Z_n\right)\\
&\le\Pi_{\mathcal{F}}(n)\max_{v\in\mathcal{F}_{|\mathbf{x}_1^n}}\Pr\left(\left|\frac{1}{n}\sum_{j=1}^{n}\sigma_j\,\mathbf{1}[v_j\ne Y_j]\right|>\varepsilon/4\mid Z_1,\ldots,Z_n\right).
\end{aligned}$$

Finally, note that conditional on $(Z_1,\ldots,Z_n)$, each random variable $\sigma_j\,\mathbf{1}[v_j\ne Y_j]$ is a zero-mean random variable taking values in $[-1,1]$. Applying Hoeffding's inequality with $[a_j,b_j]=[-1,1]$ for $j\in[n]$, the above probability is at most

$$2\Pi_{\mathcal{F}}(n)e^{-n\varepsilon^2/32}.$$

The final bound follows, even without the condition $n\varepsilon^2\ge 2$, since $8e^{-n\varepsilon^2/32}>1$ for $n\varepsilon^2<2$.

□