# Minimax lower bounds

Maxim Raginsky

December 4, 2013

Now that we have a good handle on the performance of ERM and its variants, it is time to ask whether we can do better. For example, consider binary classification: we observe $n$ i.i.d. training samples from an unknown joint distribution $P$ on $\mathsf{X} \times \{0,1\}$, where $\mathsf{X}$ is some feature space, and for a fixed class $\mathscr{F}$ of candidate classifiers $f : \mathsf{X} \to \{0,1\}$ we let $\widehat{f}_n$ be the ERM solution

$$\widehat{f}_n = \operatorname*{arg\,min}_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^{n} 1_{\{f(X_i) \neq Y_i\}}. \tag{1}$$

IF $\mathscr{F}$ is a VC class with VC dimension $V$, then the excess risk of $\widehat{f}_n$ over the best-in-class performance $L^*(\mathscr{F}) \equiv \inf_{f \in \mathscr{F}} L(f)$ satisfies

$$L(\widehat{f}_n) \leq L^*(\mathscr{F}) + C\left(\sqrt{\frac{V}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right)$$

with probability at least $1-\delta$, where $C > 0$ is some absolute constant. Integrating, we also get the following bound on the expected excess risk:

$$\mathbb{E}\left[L(\widehat{f}_n) - L^*(\mathscr{F})\right] \leq C\sqrt{\frac{V}{n}}, \tag{2}$$

for some constant $C > 0$. Crucially, the bound (2) holds for *all* possible joint distributions $P$ on $\mathsf{X} \times \{0,1\}$, and the right-hand side is *independent* of $P$ — it depends only on the properties of the class $\mathscr{F}$! Thus, we deduce the following remarkable *distribution-free guarantee* for ERM: for any VC class $\mathscr{F}$, the ERM algorithm (1) satisfies

$$\sup_{P \in \mathscr{P}(\mathsf{X} \times \{0,1\})} \mathbb{E}_P\left[L_P(\widehat{f}_n) - L_P^*(\mathscr{F})\right] \leq C\sqrt{\frac{V}{n}}. \tag{3}$$

(We have used subscript $P$ to explicitly indicate the fact that the quantity under the supremum depends on the underlying distribution $P$. In the sequel, we will often drop the subscript to keep the notation uncluttered.) Let's take a moment to reflect on the significance of the bound (3). What it says is that, regardless of how "weird" the stochastic relationship between the feature $X \in \mathsf{X}$ and the label $Y \in \{0,1\}$ might be, as long as we scale our ambition back and aim at approaching the performance of the best classifier in some VC class $\mathscr{F}$, the ERM algorithm will produce a good classifier with a uniform $O(\sqrt{V/n})$ guarantee on its excess risk.

At this point, we stop and ask ourselves: could this bound be too pessimistic, even when we are so lucky that the optimal (Bayes) classifier happens to be in $\mathscr{F}$? (Recall that the Bayes classifier for a given $P$ has the form

$$f^*(x) = \begin{cases} 1, & \text{if } \eta(x) \geq 1/2 \\ 0, & \text{otherwise} \end{cases},$$

where $\eta(x) = \mathbb{E}[Y|X=x] = \mathbb{P}(Y=1|X=x)$ is the *regression function*.) Let $\mathscr{P}(\mathscr{F})$ denote the subset of $\mathscr{P}(\mathsf{X} \times \{0,1\})$ consisting of all joint distributions of $X \in \mathsf{X}$ and $Y \in \{0,1\}$, such that $f^* \in \mathscr{F}$. Then from (2) we have

$$\sup_{P \in \mathscr{P}(\mathscr{F})} \mathbb{E}\left[L(\widehat{f}_n) - L(f^*)\right] \leq C\sqrt{\frac{V}{n}}, \tag{4}$$

where $\widehat{f}_n$ is the ERM solution (1). However, we know that if the relationship between $X$ and $Y$ is deterministic, i.e., if $Y = f^*(X)$, then ERM performs much better. More precisely, let $\mathscr{P}_0(\mathscr{F})$ be the *zero-error class*:

$$\mathscr{P}_0(\mathscr{F}) \triangleq \left\{P \in \mathscr{P}(\mathscr{F}) : Y = f^*(X) \text{ a.s.}\right\}.$$

Then one can show that

$$\sup_{P \in \mathscr{P}_0(\mathscr{F})} \mathbb{E}\left[L(\widehat{f}_n) - L(f^*)\right] \leq C\frac{V}{n}, \tag{5}$$

a much better bound than the "global" bound (4) (see, e.g., the book by Vapnik [Vap98]). This suggests that the performance of ERM is somehow tied to how "sharp" the behavior of $\eta$ is around the decision boundary that separates the sets $\{x \in \mathsf{X} : \eta(x) \geq 1/2\}$ and $\{x \in \mathsf{X} : \eta(x) < 1/2\}$. To see whether this is the case, let us define, for any $h \in [0,1]$, the class of distributions

$$\mathscr{P}(h, \mathscr{F}) \triangleq \left\{P \in \mathscr{P}(\mathscr{F}) : |2\eta(X) - 1| \geq h \text{ a.s.}\right\}$$

(in that case, the distributions in $\mathscr{P}(h,\mathscr{F})$ are said to satisfy the *Massart noise condition* with margin $h$.) We have already seen the two extreme cases:

- $h = 0$ — this gives $\mathscr{P}(0,\mathscr{F}) = \mathscr{P}(\mathscr{F})$ (the bound $|2\eta - 1| \geq 0$ holds trivially for any $P$).
- $h = 1$ — this gives the zero-error regime $\mathscr{P}(1,\mathscr{F}) = \mathscr{P}_0(\mathscr{F})$ (if $|2\eta - 1| \geq 1$ a.s., then $\eta$ can take only values 0 and 1 a.s.).

However, intermediate values of $h$ are also of interest: if a distribution $P$ belongs to $\mathscr{P}(h,\mathscr{F})$ for some $0 < h < 1$, then its regression function $\eta$ makes a jump of size $h$ as we cross the decision boundary. With this in mind, for any $n \in \mathbb{N}$ and $h \in [0,1]$ let us define the *minimax risk*

$$R_n(h,\mathscr{F}) \triangleq \inf_{\tilde{f}_n} \sup_{P \in \mathscr{P}(h,\mathscr{F})} \mathbb{E}\left[L(\tilde{f}_n) - L(f^*)\right], \tag{6}$$

where the infimum is over *all* learning algorithms $\tilde{f}_n$ based on $n$ i.i.d. training samples. The term "minimax" indicates that we are *minimizing* over all admissible learning algorithms, while *maximizing* over all distributions in a given class. The following result was proved by Pascal Massart and Élodie Nédélec [MN06]:

**Theorem 1.** *Let $\mathscr{F}$ be a VC class of binary-valued functions on $\mathsf{X}$ with VC dimension $V \geq 2$. Then for any $n \geq V$ and any $h \in [0,1]$ we have the lower bound*

$$R_n(h,\mathscr{F}) \geq c \min\left(\sqrt{\frac{V}{n}}, \frac{V}{nh}\right), \tag{7}$$

*where $c > 0$ is some absolute constant.*

Let us examine some implications:

- When $h = 0$, the right-hand side of (7) is equal to $c\sqrt{V/n}$. Thus, without any further assumptions, ERM is as good as it gets (it is minimax-optimal), up to multiplicative constants.
- When $h = 1$ (the zero-error case), the right-hand side of (7) is equal to $cV/n$, which matches the upper bound (5) up to constants. Thus, if we happen to know that we are in a zero-error situation, ERM is minimax-optimal as well.
- For intermediate values of $h$, the lower bound depends on the relative sizes of $h$, $V$, and $n$. In particular, if $h \geq \sqrt{V/n}$, we have the minimax lower bound $R_n(h,\mathscr{F}) \geq cV/nh$. Alternatively, for a fixed $h \in (0,1)$, we may think of $n^* = \lceil V/h^2 \rceil$ as the *cutoff sample size*, beyond which the effect of the margin condition on $\eta$ can be "spotted" and exploited by a learning algorithm.
- In the same paper, Massart and Nédélec obtain the following upper bound on ERM:

$$\sup_{P \in \mathscr{P}(h,\mathscr{F})} \mathbb{E}\left[L(\widehat{f}_n) - L(f^*)\right] \leq \begin{cases} C\sqrt{\dfrac{V}{n}}, & \text{if } h \leq \sqrt{V/n} \\ C\dfrac{V}{nh}\left(1 + \log\dfrac{nh^2}{V}\right), & \text{if } h > \sqrt{V/n} \end{cases}. \tag{8}$$

  Thus, ERM is nearly minimax-optimal (we say "nearly" because of the extra log factor in the above bound; in fact, as Massart and Nédélec show, the log factor is unavoidable when the function class $\mathscr{F}$ is "rich" in a certain sense). The proof of the above upper bound is rather involved and technical, and we will not get into it here.

The appearance of the logarithmic term in (8) is rather curious. Given the lower bound of Theorem 1, one may be tempted to dismiss it as an artifact of the analysis used by Massart and Nédélec. However, as we will now see, in certain situations this logarithmic term is unavoidable. To that end, we first need a definition: We say that a class $\mathscr{F}$ binary-valued functions $f : \mathsf{X} \to \{0,1\}$ is $(N,D)$-*rich*, for some $N, D \in \mathbb{N}$, if there exist $N$ distinct points $x_1, \ldots, x_N \in \mathsf{X}$, such that the projection

$$\mathscr{F}(x^N) = \left\{\big(f(x_1), \ldots, f(x_N)\big) : f \in \mathscr{F}\right\}$$

of $\mathscr{F}$ onto $x^N$ contains all binary strings of Hamming weight[1] $D$. Some examples:

- If $\mathscr{F}$ is a VC-class with VC dimension $V$, then it is $(V,D)$-rich for all $1 \leq D \leq V$. This follows directly from definitions.

[1] The Hamming weight of a binary string is the number of nonzero bits it has.

- A nontrivial example, and one that is relevant to statistical learning, is as follows. Let $\mathscr{F}$ be the collection of indicators of all halfspaces in $\mathbb{R}^d$, for some $d \geq 2$. There is a result in computational geometry which says that, for any $N \geq d+1$, one can find $N$ distinct points $x_1, \ldots, x_N$, such that $\mathscr{F}(x^n)$ contains all strings in $\{0,1\}^N$ with Hamming weight up to, and including, $\lfloor d/2 \rfloor$. Consequently, $\mathscr{F}$ is $(N, \lfloor d/2 \rfloor)$-rich for all $N \geq d+1$.

We can now state the following result [MN06]:

**Theorem 2.** *Given some $D \geq 1$, suppose that $\mathscr{F}$ is $(N, D)$-rich for all $N \geq 4D$. Then*

$$R_n(h, \mathscr{F}) \geq c(1-h)\frac{D}{nh}\left[1 + \log\frac{nh^2}{D}\right] \tag{9}$$

*for any $\sqrt{D/n} \leq h < 1$, where $c > 0$ is some absolute constant.*

We will present the proofs of Theorems 1 and 2 in Sections 2 and 3, after giving some necessary background on minimax lower bounds.

# 1 Preparation: minimax lower bounds for statistical estimation

To prove Theorem 1, we need some background on minimax lower bounds for statistical estimation problems. The presentation here follows an excellent paper by Bin Yu [Yu97].

The setting is as follows: we have an indexed set $\{P_\theta : \theta \in \Theta\}$ of probability distributions on some set $\mathsf{Z}$, where $\Theta$ is some parameter set. For our purposes, it suffices to consider the case when $\mathsf{Z}$ is a finite set. We observe a random sample $Z$ from one of the $P_\theta$'s, and our goal is to estimate $\theta$. We will measure the quality of any estimator $\widehat{\theta} = \widehat{\theta}(Z)$ by its expected risk

$$\mathbb{E}_\theta[d(\theta, \widehat{\theta}(Z))] = \sum_{z \in \mathsf{Z}} P_\theta(z) d(\theta, \widehat{\theta}(z)),$$

where $\mathbb{E}_\theta[\cdot]$ denotes expectation with respect to $P_\theta$, and $d : \Theta \times \Theta \to \mathbb{R}^+$ is a given loss function. Here, we will assume that $d$ is a *pseudometric* on $\Theta$, i.e., it has the following properties:

1. Symmetry – $d(\theta, \theta') = d(\theta', \theta)$ for any $\theta, \theta' \in \Theta$.
2. Triangle inequality – $d(\theta, \theta') \leq d(\theta, \eta) + d(\eta, \theta')$ for any $\theta, \theta', \eta \in \Theta$.

(We do not require that $d(\theta, \theta') = 0$ if and only if $\theta = \theta'$, in which case $d$ is a *metric*.) Since we do not know ahead of time which of the $\theta$'s we will be facing, it is natural to expect the worst and seek an estimator $\widehat{\theta}$ to minimize the *worst-case* risk $\sup_{\theta \in \Theta} \mathbb{E}_\theta[d(\theta, \widehat{\theta}(Z))]$. With this in mind, let us define the *minimax risk*

$$\mathfrak{M}(\Theta) \triangleq \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta[d(\theta, \widehat{\theta}(Z))],$$

where the infimum is over all estimators $\widehat{\theta} = \widehat{\theta}(Z)$. We are particularly interested in tight *lower bounds* on $\mathfrak{M}(\Theta)$, since they will give us some idea of how difficult the estimation problem is.

A simple yet powerful technique for getting lower bounds is the *two-point method* introduced by Lucien Le Cam:

**Lemma 1.** *For any two $\theta, \theta' \in \Theta$ and any estimator $\widehat{\theta}$,*

$$\mathbb{E}_\theta[d(\theta, \widehat{\theta}(Z))] + \mathbb{E}_{\theta'}[d(\theta', \widehat{\theta}(Z))] \geq d(\theta, \theta') \cdot \sum_{z \in \mathsf{Z}} \min\left(P_\theta(z), P_{\theta'}(z)\right). \tag{10}$$

*Proof.* Consider an arbitrary $z \in \mathsf{Z}$. If $P_\theta(z) \leq P_{\theta'(z)}$, then

$$\begin{aligned} P_\theta(z) d(\theta, \widehat{\theta}(z)) + P_{\theta'} d(\theta', \widehat{\theta}(z)) &= P_\theta(z) \left[ d(\theta, \widehat{\theta}(z)) + d(\theta', \widehat{\theta}(z)) \right] + \left[ P_{\theta'}(z) - P_\theta(z) \right] d(\theta', \widehat{\theta}(z)) \\ &\geq P_\theta(z) \left[ d(\theta, \widehat{\theta}(z)) + d(\theta', \widehat{\theta}(z)) \right] \\ &\geq P_\theta(z) d(\theta, \theta'), \end{aligned}$$

where the second line is because $P_{\theta'}(z) \geq P_\theta(z)$ and $d(\cdot, \cdot)$ is nonnegative, while the third line is by the triangle inequality. By the same token, if $P_\theta(z) \geq P_{\theta'}(z)$, then

$$P_\theta(z) d(\theta, \widehat{\theta}(z)) + P_{\theta'} d(\theta', \widehat{\theta}(z)) \geq P_{\theta'}(z) d(\theta, \theta').$$

Summing over all $z \in \mathsf{Z}$, we get (10). □

The sum on the right-hand side of (10) can be expressed in terms of the so-called *total variation distance* between $P_\theta$ and $P_{\theta'}$. For any two probability distributions $P, Q \in \mathscr{P}(\mathsf{Z})$, the total variation distance is

$$\|P - Q\|_{\mathrm{TV}} \triangleq \frac{1}{2} \sum_{z \in \mathsf{Z}} |P(z) - Q(z)|. \tag{11}$$

Moreover,

$$\|P - Q\|_{\mathrm{TV}} = 1 - \sum_{z \in \mathsf{Z}} \min(P(z), Q(z)). \tag{12}$$

To prove (12), we just use the definition (11): if we let $\mathsf{Z}' \triangleq \left\{ z \in \mathsf{Z} : P(z) \geq Q(z) \right\}$, then

$$\begin{aligned} \|P - Q\|_{\mathrm{TV}} &= \frac{1}{2} \sum_{z \in \mathsf{Z}'} (P(z) - Q(z)) + \frac{1}{2} \sum_{z \notin \mathsf{Z}'} (Q(z) - P(z)) \\ &= \frac{1}{2} \left( P(\mathsf{Z}') - Q(\mathsf{Z}') \right) + \frac{1}{2} \left( Q(\mathsf{Z}'^c) - P(\mathsf{Z}'^c) \right) \\ &= \frac{1}{2} \left( P(\mathsf{Z}') - Q(\mathsf{Z}') \right) + \frac{1}{2} \left( P(\mathsf{Z}') - Q(\mathsf{Z}') \right) \\ &= P(\mathsf{Z}') - Q(\mathsf{Z}') \\ &= \sum_{z : P(z) \geq Q(z)} (P(z) - Q(z)) \\ &= \sum_{z : P(z) \geq Q(z)} (P(z) - \min(P(z), Q(z))) \\ &= \sum_{z \in \mathsf{Z}} (P(z) - \min(P(z), Q(z))) - \underbrace{\sum_{z : P(z) \leq Q(z)} (P(z) - \min(P(z), Q(z)))}_{=0} \\ &= 1 - \sum_{z \in \mathsf{Z}} \min(P(z), Q(z)). \end{aligned}$$

and we are done. Thus, we can rewrite the bound (10) as

$$\mathbb{E}_\theta[d(\theta, \widehat{\theta}(Z))] + \mathbb{E}_{\theta'}[d(\theta', \widehat{\theta}(Z))] \geq d(\theta, \theta') \cdot \left(1 - \|P_\theta - P_{\theta'}\|_{\mathrm{TV}}\right). \tag{13}$$

Let us examine some consequences. If we take the supremum of both sides of (13) over the choices of $\theta$ and $\theta'$, then we see that, for any estimator $\widehat{\theta}$,

$$\sup_{\theta\in\Theta}\mathbb{E}_\theta[d(\theta,\widehat{\theta}(Z))] \geq \frac{1}{2}\sup_{\theta\neq\theta'}\left\{d(\theta,\theta')\cdot\left(1-\|P_\theta - P_{\theta'}\|_{\mathrm{TV}}\right)\right\}. \tag{14}$$

Notice that the right-hand side does not depend on the choice of $\widehat{\theta}$, so we can take the infimum of both sides of (14) over $\widehat{\theta}$ and obtain the following:

**Corollary 1.** *The minimax risk $\mathfrak{M}(\Theta)$ can be lower-bounded as*

$$\mathfrak{M}(\Theta) \geq \frac{1}{2}\sup_{\theta\neq\theta'}\left\{d(\theta,\theta')\cdot\left(1-\|P_\theta - P_{\theta'}\|_{\mathrm{TV}}\right)\right\}. \tag{15}$$

The "two-point" bound (15) gives us an idea of when the problem of estimating $\theta$ based on observations $Z\sim P_\theta$ is difficult: when there exists at least one pair of distinct parameter values $\theta,\theta'\in\Theta$ such that $d(\theta,\theta')$ is large, while the total variation distance $\|P_\theta - P_{\theta'}\|_{\mathrm{TV}}$ is small. This means that the observation $Z$ does not give us enough information to reliably distinguish between $\theta$ and $\theta'$, but the consequences of mistaking one for the other are severe (since $d(\theta,\theta')$ is large).

Another useful consequence of Lemma 1 comes about when we consider the Bayesian setting, i.e., when we have a *prior distribution* $\pi$ on $\Theta$ and consider the *average risk* of any estimator $\widehat{\theta}$:

$$\mathbb{E}_\pi[d(\theta,\widehat{\theta}(Z))] \triangleq \int_\Theta \pi(d\theta)\mathbb{E}_\theta[d(\theta,\widehat{\theta}(Z))].$$

Then we have the following lower bound on $\mathfrak{M}(\Theta)$:

$$\mathfrak{M}(\Theta) \geq \inf_{\widehat{\theta}}\mathbb{E}_\pi\left[d(\theta,\widehat{\theta}(Z))\right].$$

To prove this, we simply note that, for any $\widehat{\theta}$,

$$\mathbb{E}_\pi[d(\theta,\widehat{\theta}(Z))] \triangleq \int_\Theta \pi(d\theta)\mathbb{E}_\theta[d(\theta,\widehat{\theta}(Z))] \leq \sup_{\theta\in\Theta}\mathbb{E}_\theta[d(\theta,\widehat{\theta})].$$

(In fact, under suitable regularity conditions, one can prove that the minimax risk is equal to the *worst-case Bayes risk*:

$$\mathfrak{M}(\Theta) = \inf_{\widehat{\theta}}\sup_{\pi\in\mathscr{P}(\Theta)}\mathbb{E}_\pi\left[d(\theta,\widehat{\theta}(Z))\right] = \sup_{\pi\in\mathscr{P}(\Theta)}\inf_{\widehat{\theta}}\mathbb{E}_\pi\left[d(\theta,\widehat{\theta}(Z))\right],$$

where the first equality always holds, while the second equality is valid whenever the conditions of the minimax theorem are satisfied.) With these definitions, we have the following:[2]

**Corollary 2.** *Let $\pi$ be any prior distribution on $\Theta$, and let $\mu$ be any joint probability distribution of a random pair $(\theta,\theta')\in\Theta\times\Theta$, such that the marginal distributions of both $\theta$ and $\theta'$ are equal to $\pi$. Then*

$$\mathfrak{M}(\Theta) \geq \inf_{\widehat{\theta}}\mathbb{E}_\pi\left[d(\theta,\widehat{\theta}(Z))\right] \geq \mathbb{E}_\mu\left[d(\theta,\theta')\cdot\left(1-\|P_\theta - P_{\theta'}\|_{\mathrm{TV}}\right)\right]. \tag{16}$$

[2] I have learned this particular formulation from my colleagues Yihong Wu (here at Illinois) and Yury Polyanskiy (MIT).

*Proof.* Take the expectation of both sides of (13) with respect to $\mu$ and then use the fact that, under $\mu$, both $\theta$ and $\theta'$ have the same distribution $\pi$. □

In many cases, the analysis of minimax lower bounds can be reduced to the following problem: Suppose that the parameter set $\Theta$ can be identified with the $m$-dimensional binary hypercube $\{0,1\}^m$ for some $m$, and the objective is to determine every bit of the underlying unknown parameter $\theta \in \{0,1\}^m$ based on an observation $Z \sim P_\theta$. In that setting, a key result known as *Assouad's lemma* says that the difficulty of estimating the *entire* bit string $\theta$ is related to the difficulty of estimating each bit of $\theta$ separately, assuming you already know all other bits:

**Theorem 3** (Assouad's lemma)**.** *Suppose that* $\Theta = \{0,1\}^m$ *and consider the Hamming metric*

$$d_{\mathrm{H}}(\theta,\theta') \triangleq \sum_{i=1}^{m} 1_{\{\theta_i \neq \theta'_i\}}. \tag{17}$$

*Then*

$$\mathfrak{M}(\Theta) \geq \frac{m}{2}\left(1 - \max_{\theta,\theta' : d_{\mathrm{H}}(\theta,\theta')=1} \|P_\theta - P_{\theta'}\|_{\mathrm{TV}}\right). \tag{18}$$

*Proof.* Notice that we can write $d_{\mathrm{H}}(\theta,\theta')$ as a sum $\sum_{i=1}^m d_i(\theta,\theta')$, where $d_i(\theta,\theta') \triangleq 1_{\{\theta_i \neq \theta'_i\}}$, and each $d_i$ is a pseudometric. Now let $\pi$ be the uniform distribution on $\Theta = \{0,1\}^m$, and for each $i \in \{1,\ldots,m\}$ let $\mu_i$ be the joint distribution of a random pair $(\theta,\theta') \in \Theta \times \Theta$, under which $\theta \sim \pi$ and $\theta'$ differs from $\theta$ only in the $i$th bit, i.e., $\theta \sim \pi$ and $d_i(\theta,\theta') = 1$. Then the marginal distribution of $\theta'$ under $\mu_i$ is

$$\sum_{\theta \in \{0,1\}^m} \mu_i(\theta,\theta') = \frac{1}{2^m} \sum_{\theta \in \{0,1\}^m} 1_{\{\theta_i \neq \theta'_i \text{ and } \theta_j = \theta'_j, j \neq i\}} = \frac{1}{2^m} = \pi(\theta'),$$

since for each $\theta'$ there is only one $\theta$ that differs from it in a single bit. Now the idea is to apply Corollary 2 separately to each bit:

$$\begin{aligned}
\mathfrak{M}(\Theta) &= \inf_{\widehat{\theta}} \mathbb{E}_\pi\left[d_{\mathrm{H}}(\theta,\widehat{\theta}(Z))\right] \\
&= \inf_{\widehat{\theta}} \sum_{i=1}^m \mathbb{E}_\pi\left[d_i(\theta,\widehat{\theta}(Z))\right] \\
&\geq \sum_{i=1}^m \inf_{\widehat{\theta}} \mathbb{E}_\pi\left[d_i(\theta,\widehat{\theta}(Z))\right] \\
&\geq \sum_{i=1}^m \frac{1}{2}\mathbb{E}_{\mu_i}\left[d_i(\theta,\theta') \cdot \left(1 - \|P_\theta - P_{\theta'}\|_{\mathrm{TV}}\right)\right],
\end{aligned}$$

where the last step is by Corollary 2. Since $d_i(\theta,\theta') = 1$ under $\mu_i$ for every $i$, we have

$$\begin{aligned}
\mathfrak{M}(\Theta) &\geq \frac{1}{2}\sum_{i=1}^m \mathbb{E}_{\mu_i}\left[1 - \|P_\theta - P_{\theta'}\|_{\mathrm{TV}}\right] \\
&\geq \frac{1}{2}\sum_{i=1}^m \min_{\theta,\theta' : d_{\mathrm{H}}(\theta,\theta')=1} \left(1 - \|P_\theta - P_{\theta'}\|_{\mathrm{TV}}\right) \\
&= \frac{m}{2}\left(1 - \max_{\theta,\theta' : d_{\mathrm{H}}(\theta,\theta')=1} \|P_\theta - P_{\theta'}\|_{\mathrm{TV}}\right),
\end{aligned}$$

and we are done. □

In order to apply Assouad's lemma, we need to have an upper bound on the total variation distance between any two $P_\theta$ and $P_{\theta'}$ with $d_{\mathrm{H}}(\theta,\theta') = 1$. More often than not, it is easier to work with another distance between probability distributions, the so-called *Hellinger distance*. For any two $P, Q \in \mathscr{P}(\mathsf{Z})$, their *squared* Hellinger distance is given by

$$H^2(P,Q) \triangleq \frac{1}{2}\sum_{z\in\mathsf{Z}}\left(\sqrt{P(z)} - \sqrt{Q(z)}\right)^2.$$

The total variation distance can be both upper- and lower-bounded by Hellinger:

$$\frac{1}{2}H^2(P,Q) \le \|P-Q\|_{\mathrm{TV}} \le H(P,Q). \tag{19}$$

Moreover, for any $n$ pairs of distributions $(P_1,Q_1),\ldots,(P_n,Q_n)$,

$$H^2(P_1\times\ldots\times P_n, Q_1\times\ldots\times Q_n) \le \sum_{i=1}^n H^2(P_i,Q_i). \tag{20}$$

Armed with these facts, we can prove the following version of Assouad's lemma:

**Corollary 3.** *Let $\{P_\theta : \theta \in \{0,1\}^m\}$ be a collection of probability distributions on some set $\mathsf{Z}$ indexed by the vertices of the binary hypercube $\Theta = \{0,1\}^m$. Suppose that there exists some constant $\alpha > 0$, such that*

$$H^2(P_\theta, P_{\theta'}) \le \alpha, \qquad \text{if } d_{\mathrm{H}}(\theta,\theta') = 1. \tag{21}$$

*Consider the problem of estimating the parameter $\theta \in \Theta$ based on $n$ i.i.d. observations from $P_\theta$, where the loss is measured by the Hamming distance. Then the corresponding minimax risk, which we denote by $\mathfrak{M}_n(\Theta)$, is lower-bounded as*

$$\mathfrak{M}_n(\Theta) \ge \frac{m}{2}(1-\sqrt{\alpha n}). \tag{22}$$

*Proof.* For any $\theta \in \Theta$, let $P_\theta^n$ denote the product of $n$ copies of $P_\theta$, i.e., the joint distribution of $n$ i.i.d. samples from $P_\theta$. For any two $\theta,\theta' \in \Theta$ with $d_{\mathrm{H}}(\theta,\theta') = 1$, we have

$$\begin{aligned}
\|P_\theta^n - P_{\theta'}^n\|_{\mathrm{TV}} &\le H(P_\theta^n, Q_\theta^n) \\
&\le \sqrt{\sum_{i=1}^n H^2(P_\theta, P_{\theta'})} \\
&\le \sqrt{\alpha n},
\end{aligned}$$

where the first step is by (19), the second is by (20), and the third is by (21). The bound (22) then follows from Assouad's lemma. □

Assouad's lemma is a wonderful tool, but it is only applicable when the parameter space $\Theta$ is a binary hypercube and the metric $d$ is the Hamming metric. To handle other situations, we need different methods. Let's recall what we had seen earlier: the minimax risk $\mathfrak{M}(\Theta)$ is large if we can find at least two distinct points $\theta,\theta' \in \Theta$, such that (a) the distance $d(\theta,\theta')$ is large and (b) the distributions $P_\theta$ and $P_{\theta'}$ are statistically close to one another, i.e., the total variation distance $\|P_\theta - P_{\theta'}\|_{\mathrm{TV}}$ is small. Let's see what happens when we follow this logic further and consider any two $\theta,\theta' \in \Theta$ that are separated from each other by a distance of at least $\varepsilon > 0$:

**Lemma 2.** *Fix $\varepsilon > 0$, and let $\Theta_0 \subset \Theta$ be any set which is $\varepsilon$-separated, i.e., $d(\theta,\theta') \geq \varepsilon$ for any two distinct $\theta,\theta' \in \Theta_0$. Then*

$$\mathfrak{M}(\Theta) \geq \frac{\varepsilon}{2} \inf_{\tilde{\theta}} \sup_{\theta \in \Theta_0} \mathbb{P}_\theta\left(\tilde{\theta} \neq \theta\right), \tag{23}$$

*where the infimum is over all estimators $\tilde{\theta}$ that take values in $\Theta_0$.*

*Proof.* Fix any $\varepsilon$-separated set $\Theta_0$. Then, since $\Theta_0 \subset \Theta$, we have $\mathfrak{M}(\Theta) \geq \mathfrak{M}(\Theta_0)$. Given any estimator $\widehat{\theta} = \widehat{\theta}(Z)$ taking values in $\Theta$, define another estimator $\tilde{\theta} = \tilde{\theta}(Z)$ taking values in $\Theta_0$ as follows:

$$\tilde{\theta} = \operatorname*{arg\,min}_{\theta' \in \Theta_0} d(\widehat{\theta}, \theta').$$

Then, for any $\theta \in \Theta_0$,

$$d(\tilde{\theta},\theta) \leq d(\tilde{\theta},\widehat{\theta}) + d(\theta,\widehat{\theta}) \leq 2d(\theta,\widehat{\theta}),$$

where the first step is by the triangle inequality, while the second step is by definition of $\tilde{\theta}$. Consequently,

$$\mathfrak{M}(\Theta) \geq \mathfrak{M}(\Theta_0) \geq \frac{1}{2} \inf_{\tilde{\theta}} \sup_{\theta \in \Theta_0} \mathbb{E}_\theta[d(\tilde{\theta},\theta)]. \tag{24}$$

Now let us consider the event $\{z \in \mathsf{Z} : \tilde{\theta}(z) \neq \theta\}$ for some $\theta \in \Theta_0$. Since $\tilde{\theta} \in \Theta_0$ by construction, and $\Theta_0$ is $\varepsilon$-separated, we have $\{z \in \mathsf{Z} : \tilde{\theta}(z) \neq \theta\} \subseteq \{z \in \mathsf{Z} : d(\tilde{\theta}(z),\theta) \geq \varepsilon\}$. Using this fact and Markov's inequality, we can write

$$\mathbb{P}_\theta\left(\tilde{\theta} \neq \theta\right) \leq \mathbb{P}_\theta\left(d(\tilde{\theta},\theta) \geq \varepsilon\right) \leq \frac{\mathbb{E}_\theta\left[d(\tilde{\theta},\theta)\right]}{\varepsilon}. \tag{25}$$

Using this in (24), we get (23). □

Lemma 2 is at its most effective when we can find a *finite* set $\Theta_0$ which is $\varepsilon$-separated and has large cardinality. The reason for the latter requirement is that then it should not be too hard to arrange things in such a way that the probability of error $\mathbb{P}_\theta(\tilde{\theta} \neq \theta)$ is uniformly bounded away from zero for *any* estimator $\tilde{\theta}$. In particular, if we can find an $\varepsilon$-separated set $\Theta_0 \subset \Theta$, such that

$$\inf_{\tilde{\theta}} \max_{\theta \in \Theta_0} \mathbb{P}_\theta(\tilde{\theta},\theta) \geq \alpha$$

for some absolute constant $\alpha > 0$, then we can lower-bound the minimax risk $\mathfrak{M}(\Theta)$ by $\alpha\varepsilon/2$. So now we need tight lower bounds on the probability of error when testing multiple hypotheses that hold for *any* estimation procedure. Bounds of this sort rely, in one way or another, on a fundamental result from information theory known as *Fano's inequality* [CT06] (or Fano's lemma, as statisticians call it [Yu97]). We will use a particular version, due to Lucien Birgé [Bir05]. To state it, we first need to define relative entropy (or Kullback–Leibler divergence). Given any two probability distributions $P, Q$ on $\mathsf{Z}$, the *relative entropy* (or *Kullback–Leibler divergence*) between $P$ and $Q$ is defined as

$$D(P\|Q) \triangleq \begin{cases} \displaystyle\sum_{z \in \mathsf{Z}} P(z) \log \frac{P(z)}{Q(z)}, & \text{if } \operatorname{supp}(P) \subseteq \operatorname{supp}(Q) \\ +\infty, & \text{otherwise} \end{cases} \tag{26}$$

where $\operatorname{supp}(P) \triangleq \{z \in \mathsf{Z} : P(z) > 0\}$ is the support of $P$. Here are some useful properties:

- $D(P\|Q) \geq 0$, with equality if and only if $P \equiv Q$;
- for any $n$ pairs of distributions $(P_i, Q_i)$, $1 \leq i \leq n$, we have

$$D(P_1 \times \ldots \times P_n \| Q_1 \times \ldots \times Q_n) = \sum_{i=1}^{n} D(P_i \| Q_i). \tag{27}$$

Also, one can bound the total variation distance in terms of the KL divergence as follows:

$$\|P - Q\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{2} D(P\|Q)}. \tag{28}$$

This inequality is usually referred to as *Pinsker's inequality* after Mark S. Pinsker, although Pinsker didn't prove it. In the present form, the inequality (28), with the optimal constant $1/2$ in front of $D(P\|Q)$, was obtained independently by I. Csiszár, J.H.B. Kemperman, and S. Kullback.

With these preliminaries out of the way, we can state Birgé's bound:

**Lemma 3** (Birgé)**.** *Let $\{P_i\}_{i=0}^N$ be a finite family of probability distributions on $\mathsf{Z}$, and let $\{A_i\}_{i=0}^N$ be a family of disjoint events. Then*

$$\min_{0 \leq i \leq N} P_i(A_i) \leq \max\left(\alpha, \frac{\bar{K}}{\log(N+1)}\right),$$

*where $\alpha = 0.71$ and*

$$\bar{K} = \frac{1}{N} \sum_{i=1}^{N} D(P_i \| P_0).$$

Now let's see how we can apply Birgé's lemma: Let $\Theta_0 = \{\theta_0, \ldots, \theta_N\}$ be some $\varepsilon$-separated subset of $\Theta$. Fix an estimator $\tilde{\theta}$ taking values in $\Theta_0$. For each $0 \leq i \leq N$, consider the probability distribution $P_i = P_{\theta_i}$ and the event $A_i = \{z \in \mathsf{Z} : \tilde{\theta}(z) = \theta_i\}$. Since the elements of $\Theta_0$ are all distinct, the events $A_0, \ldots, A_N$ are disjoint. Now suppose that we have chosen $\Theta_0$ in such a way that

$$\bar{K} = \frac{1}{N} \sum_{i=1}^{N} D(P_0 \| P_i) = \frac{1}{N} \sum_{i=1}^{N} D\big(P_{\theta_0} \| P_{\theta_i}\big) \leq \alpha \log(N+1). \tag{29}$$

Then, by Birgé's lemma,

$$\begin{aligned} \max_{\theta \in \Theta_0} \mathbb{P}_\theta(\tilde{\theta} \neq \theta) &= \max_{0 \leq i \leq N} P_i(A_i^c) \\ &= 1 - \min_{0 \leq i \leq N} P_i(A_i) \\ &\geq 1 - \alpha \\ &\equiv 0.29. \end{aligned}$$

This bound holds for *any* estimator $\tilde{\theta}$. Consequently, using Lemma 2, we get the lower bound $\mathfrak{M}(\Theta) \geq (1-\alpha)\varepsilon/2$.

## 2 Proof of Theorem 1

Roughly speaking, the strategy of the proof will be as follows: We start by observing that we can lower-bound the minimax risk $R_n(h,\mathscr{F})$ by

$$R_n(h,\mathscr{F}) \geq \inf_{\tilde{f}_n} \sup_{P \in \mathscr{Q}} \mathbb{E}\left[L(\tilde{f}_n) - L(f^*)\right]$$

for any subset $\mathscr{Q} \subset \mathscr{P}(h,\mathscr{F})$. We will then construct $\mathscr{Q}$ in such a way that its elements can be naturally indexed by the vertices of a binary hypercube of dimension $V-1$ (where $V$ is the VC dimension of $\mathscr{F}$), and the expected excess risk for each $P \in \mathscr{Q}$ can be related to the minimax risk for the problem of estimating the corresponding binary string of length $V-1$. At that point, we will be in a position to apply Assouad's lemma.

### 2.1 Construction of $\mathscr{Q}$

As stated above, our set $\mathscr{Q}$ of "difficult" distributions will consist of $2^{V-1}$ distributions $P_b$, $b \in \{0,1\}^{V-1}$. To construct these distributions, we will first pick the marginal distribution $P_X \in \mathscr{P}(\mathsf{X})$ of the feature $X$, and then specify the conditional distributions $P_{Y|X}^{(b)}$ of the binary label $Y$ given $X$, for each $b \in \{0,1\}^{V-1}$. For now, let us assume that $h > 0$.

We construct $P_X$ as follows. Since $\mathscr{F}$ is a VC class with VC dimension $V$, there exists a set $\{x_1,\ldots,x_V\} \subset \mathsf{X}$ of points that can be shattered by $\mathscr{F}$, i.e., for any binary string $\beta \in \{0,1\}^V$ there exists at least one $f \in \mathscr{F}$, such that $f(x_j) = \beta_j$ for all $1 \leq j \leq V$. Given a parameter $p \in [0, 1/(V-1)]$, which will be chosen later, we choose $P_X$ so that $P_X(\{x_1,\ldots,x_V\}) = 1$, and

$$P_X(x_j) = \begin{cases} p, & \text{if } 1 \leq j \leq V-1 \\ 1-(V-1)p, & \text{otherwise} \end{cases}. \tag{30}$$

In other words, $P_X$ is a discrete distribution supported on the shattered set $\{x_1,\ldots,x_V\}$ that puts the same mass $p$ on each of the first $V-1$ points and dumps the rest of the probability mass, equal to $1-(V-1)p$, on the last point $x_V$. Owing to our condition on $p$, this is a valid probability distribution.

Next we choose $P_{Y|X}^{(b)}$ for each $b \in \{0,1\}^{V-1}$ as follows: For a fixed $b$, the conditional distribution of $Y$ given $X = x$ is

$$\begin{array}{ll} \text{Bernoulli}\left(\dfrac{1+(2b_j-1)h}{2}\right), & \text{if } x = x_j \text{ and } j \in \{1,\ldots,V-1\} \\ \text{Bernoulli}(0), & \text{otherwise} \end{array}$$

In other words, for a fixed $b$, we have

$$\eta_b(x) \equiv P_{Y|X}^{(b)}(Y=1|X=x) = \begin{cases} \frac{1-h}{2}, & \text{if } x = x_j \text{ for some } j \in \{1,\ldots,V-1\}, \text{ and } b_j = 0 \\ \frac{1+h}{2}, & \text{if } x = x_j \text{ for some } j \in \{1,\ldots,V-1\}, \text{ and } b_j = 1 \\ 0, & \text{otherwise} \end{cases}. \tag{31}$$

Therefore, the corresponding Bayes classifier, which we denote by $f_b^*$, is given by

$$f_b^*(x) = \begin{cases} 0, & \text{if } x = x_j \text{ for some } j \in \{1,\ldots,V-1\}, \text{ and } b_j = 0 \\ 1, & \text{if } x = x_j \text{ for some } j \in \{1,\ldots,V-1\}, \text{ and } b_j = 1 \\ 0, & \text{otherwise} \end{cases}. \tag{32}$$

That is, the output of the Bayes classifier on each $x_j$, $1 \le j \le V-1$, is simply equal to the bit value $b_j$, and is zero everywhere else.

It remains to check whether the resulting set $\mathscr{Q} = \left\{P_b : b \in \{0,1\}^{V-1}\right\}$ is contained in $\mathscr{P}(h,\mathscr{F})$. First of all, from (31) we see that $|2\eta_b(x)-1| \ge h$ for all $x$ (indeed, $|2\eta_b(x)-1| = h$ when $x \in \{x_1,\ldots,x_{V-1}\}$, and $|2\eta_b(x)-1| = 1$ otherwise). Secondly, because $\{x_1,\ldots,x_V\}$ is shattered by $\mathscr{F}$, there exists at least one $f \in \mathscr{F}$, such that $f_b^*(x) = f(x)$ for all $x \in \{x_1,\ldots,x_V\}$. Thus, $\mathscr{Q} \subset \mathscr{P}(h,\mathscr{F})$, as claimed.

## 2.2 Reduction to an estimation problem on the binary hypercube

Now that we have constructed $\mathscr{Q}$, we will show that the problem of learning a classifier when the $n$ training samples are drawn i.i.d. from some $P_b \in \mathscr{Q}$ is at least as difficult as reconstructing the underlying bit string $b$ — indeed, intuitively this makes sense, given the structure of the Bayes classifier $f_b^*$ corresponding to $P_b$.

We start by noting that, for any classifier $f : \mathsf{X} \to \{0,1\}$ and any distribution $P$ on $\mathsf{X} \times \{0,1\}$, we have[3]

$$L(f) - L(f^*) = \mathbb{E}\left[|2\eta(X)-1|\,|f(X)-f^*(X)|\right].$$

From this, we see that if $P \in \mathscr{P}(h,\mathscr{F})$, then

$$L(f) - L(f^*) \ge h\,\mathbb{E}\left[|f(X)-f^*(X)|\right] \equiv h\,\|f - f^*\|_{L_1},$$

where the $L_1$ norm is computed w.r.t. the marginal distribution $P_X$. Hence, we have

$$\begin{aligned}\inf_{\tilde{f}_n}\sup_{P\in\mathscr{Q}} \mathbb{E}\left[L(\tilde{f}_n) - L(f^*)\right] &= \inf_{\tilde{f}_n}\max_{b\in\{0,1\}^{V-1}} \mathbb{E}_b\left[L(\tilde{f}_n) - L(f_b^*)\right] \\ &\ge h \inf_{\tilde{f}_n}\max_{b\in\{0,1\}^{V-1}} \mathbb{E}_b\|\tilde{f}_n - f_b^*\|_{L_1}, \end{aligned} \tag{33}$$

where the $L_1$ norm is w.r.t. the distribution $P_X$ defined in (30), and $\mathbb{E}_b[\cdot]$ denotes expectation w.r.t. $P_b$. Now we are ready to carry out the reduction to an estimation problem on the binary hypercube $\{0,1\}^{V-1}$. Given any candidate estimator $\tilde{f}_n$, let $\tilde{b}_n$ be an estimator that takes values in $\{0,1\}^{V-1}$, and is defined as follows:

$$\tilde{b}_n \triangleq \operatorname*{arg\,min}_{b\in\{0,1\}^{V-1}} \|\tilde{f}_n - f_b^*\|_{L_1}. \tag{34}$$

In other words, $\tilde{b}_n$ is the binary string that indexes the element of $\left\{f_b^* : b \in \{0,1\}^{V-1}\right\}$ which is the closest to $\tilde{f}_n$ in the $L_1$ norm. Then, for any $b$,

$$\begin{aligned}\|f_{\tilde{b}_n}^* - f_b^*\|_{L_1} &\le \|f_{\tilde{b}_n}^* - \tilde{f}_n\|_{L_1} + \|\tilde{f}_n - f_b^*\|_1 \\ &\le 2\|\tilde{f}_n - f_b^*\|_{L_1}, \end{aligned} \tag{35}$$

where the first step is by the triangle inequality, while the second step is by (34). Combining (35) with (33), we get

$$\inf_{\tilde{f}_n}\sup_{P\in\mathscr{Q}} \mathbb{E}\left[L(\tilde{f}_n) - L(f^*)\right] \ge \frac{h}{2}\inf_{\tilde{b}_n}\max_{b\in\{0,1\}^{V-1}} \mathbb{E}_b\|f_{\tilde{b}_n}^* - f_b^*\|_{L_1}, \tag{36}$$

[3] Exercise: prove this!

where the infimum is over all estimators that take values in $\{0,1\}^{V-1}$ based on $n$ i.i.d. samples from one of the $P_b$'s.

Now let us inspect the $L_1$ norm $\|f_b^* - f_{b'}^*\|_{L_1}$ for any two $b, b'$. Using (32), we have

$$\begin{aligned}
\|f_b^* - f_{b'}^*\|_{L_1} &= \int_{\mathsf{X}} |f_b^*(x) - f_{b'}^*(x)| P_X(dx) \\
&= \sum_{j=1}^{V} P_X(x_j) \left| f_b^*(x_j) - f_{b'}^*(x_j) \right| \\
&= p \sum_{j=1}^{V-1} \left| f_b^*(x_j) - f_{b'}^*(x_j) \right| \\
&= p \sum_{j=1}^{V-1} \left| b_j - b'_j \right| \\
&= p\, d_{\mathrm{H}}(b, b').
\end{aligned}$$

Substituting this into (36) gives

$$\inf_{\tilde{f}_n} \sup_{P \in \mathcal{Q}} \mathbb{E}\left[ L(\tilde{f}_n) - L(f^*) \right] \geq \frac{ph}{2} \inf_{\widehat{b}_n} \max_{b \in \{0,1\}^{V-1}} \mathbb{E}_b \left[ d_{\mathrm{H}}(\widehat{b}_n, b) \right], \tag{37}$$

as claimed. Now we are in a position to apply Assouad's lemma.

### 2.3 Applying Assouad's lemma

In order to apply Assouad's lemma, we need an upper bound on the squared Hellinger distance $H^2(P_b, P_{b'})$ for all $b, b'$ with $d_{\mathrm{H}}(b, b') = 1$. This is a simple computation. In fact, for any two $b, b' \in \{0,1\}^{V-1}$ we have

$$\begin{aligned}
H^2(P_b, P_{b'}) &= \sum_{j=1}^{V} \sum_{y \in \{0,1\}} \left( \sqrt{P_b(x_j, y)} - \sqrt{P_{b'}(x_j, y)} \right)^2 \\
&= p \sum_{j=1}^{V-1} \sum_{y \in \{0,1\}} \left( \sqrt{P_{Y|X}^{(b)}(y|x_j)} - \sqrt{P_{Y|X}^{(b')}(y|x_j)} \right)^2 \\
&= p \sum_{j=1}^{V-1} H^2 \left( \mathrm{Bernoulli}\left( \frac{1 + (2b_j - 1)h}{2} \right), \mathrm{Bernoulli}\left( \frac{1 + (2b'_j - 1)h}{2} \right) \right).
\end{aligned}$$

For each $j \in \{1, \ldots, V-1\}$, the $j$th term in the above summation is nonzero if and only if $b_j \neq b'_j$, in which case it is equal to the squared Hellinger distance between the Bernoulli($\frac{1-h}{2}$) and Bernoulli($\frac{1+h}{2}$) distributions. Thus,

$$\begin{aligned}
H^2(P_b, P_{b'}) &= pH^2 \left( \mathrm{Bernoulli}\left( \frac{1-h}{2} \right), \mathrm{Bernoulli}\left( \frac{1+h}{2} \right) \right) \cdot d_{\mathrm{H}}(b, b') \\
&= 2p \left( \sqrt{\frac{1-h}{2}} - \sqrt{\frac{1+h}{2}} \right)^2 d_{\mathrm{H}}(b, b') \\
&= 2p \left( 1 - \sqrt{1-h^2} \right) d_{\mathrm{H}}(b, b').
\end{aligned}$$

In particular, the collection of distributions $\{P_b : b \in \{0,1\}^{V-1}\}$ satisfies the condition (21) of Corollary 3 with $\alpha = 2p(1-\sqrt{1-h^2}) \le 2ph^2$. Therefore, applying the bound (22), we get

$$\inf_{\tilde{f}_n} \sup_{P \in \mathscr{Q}} \mathbb{E}\left[L(\tilde{f}_n) - L(f^*)\right] \ge \frac{p(V-1)h}{4}\left(1 - \sqrt{2nph^2}\right).$$

If we let $p = 2/9nh^2$, the term in parentheses will be equal to $1/3$, and

$$\inf_{\tilde{f}_n} \sup_{P \in \mathscr{Q}} \mathbb{E}\left[L(\tilde{f}_n) - L(f^*)\right] \ge \frac{V-1}{54nh},$$

assuming that the condition $p \le 1/(V-1)$ holds. This will be the case if $h \ge \sqrt{(V-1)/n}$. Therefore,

$$\inf_{\tilde{f}_n} \sup_{P \in \mathscr{P}(h,\mathscr{F})} \mathbb{E}\left[L(\tilde{f}_n) - L(f^*)\right] \ge \frac{V-1}{54nh}, \qquad \text{if } h \ge \sqrt{(V-1)/n}. \tag{38}$$

If $h \le \sqrt{(V-1)/n}$, we can use the above construction with $\tilde{h} = \sqrt{(V-1)/n}$. In that case, because $\mathscr{P}(h,\mathscr{F}) \subseteq \mathscr{P}(h',\mathscr{F})$ whenever $h \ge h'$, we see that

$$\begin{aligned}
\inf_{\tilde{f}_n} \sup_{P \in \mathscr{P}(h,\mathscr{F})} \mathbb{E}\left[L(\tilde{f}_n) - L(f^*)\right] &\ge \inf_{\tilde{f}_n} \sup_{P \in \mathscr{P}(\tilde{h},\mathscr{F})} \mathbb{E}\left[L(\tilde{f}_n) - L(f^*)\right] \\
&\ge \frac{V-1}{54n\tilde{h}} \\
&= \frac{1}{54}\sqrt{\frac{V-1}{n}}, \qquad \text{if } h \le \sqrt{(V-1)/n}.
\end{aligned} \tag{39}$$

Putting together (38) and (39), we get the bound of Theorem 1.

## 3 Proof of Theorem 2

In its broad outline, the proof is very similar to the proof of Theorem 1. Fix some $N \ge 4D$. Since $\mathscr{F}$ is $(N,D)$-rich, there exist $N$ distinct points $x_1,\ldots,x_N \in \mathsf{X}$, such that

$$\{0,1\}_D^N \triangleq \left\{b \in \{0,1\}^N : \mathrm{wt}(b) = D\right\} \subseteq \mathscr{F}(x^n),$$

where $\mathrm{wt}(b)$ denotes the Hamming weight of a binary string $b$. Let $P_X$ be the uniform distribution on $\{x_1,\ldots,x_N\}$. Also, for each $b \in \{0,1\}_D^N$, let

$$\eta_b(x_i) \triangleq \frac{1 + (2b_i - 1)h}{2}, \qquad 1 \le i \le N.$$

Now for each such $b$ define a distribution $P_b \in \mathscr{P}(\mathsf{X} \times \{0,1\})$ by $P_b = P_X \times P_{Y|X}^{(b)}$, where

$$P_{Y|X}^{(b)}(1|x_i) = \frac{1 + (2b_i - 1)h}{2} \equiv \eta_b(x_i).$$

In other words, the conditional distribution $P_{Y|X=x_i}^{(b)}$ is a Bernoulli measure with bias $\frac{1+h}{2}$ if $b_i = 1$ and $\frac{1-h}{2}$ if $b_i = 0$. The corresponding Bayes classifier is given by $f_b^*(x_i) = b_i$ for each $i$, as before. Moreover,

since $\mathscr{F}$ is $(N,D)$-rich, for any $b \in \{0,1\}^N_D$ there exists at least one $f \in \mathscr{F}$, such that $f(x_i) = b_i = f^*_b(x_i)$ for every $i$. Consequently,

$$\mathscr{Q} = \left\{P_b : b \in \{0,1\}^N_D\right\} \subset \mathscr{P}(h,\mathscr{F}).$$

Proceeding in the same way as in the proof of Theorem 1, for any subset $\mathscr{C} \subseteq \{0,1\}^N_D$,

$$\begin{aligned} R_n(h,\mathscr{F}) &\geq \frac{h}{2} \inf_{\tilde{b}_n \in \mathscr{C}} \max_{b \in \mathscr{C}} \mathbb{E}_b \left\| f^*_{\tilde{b}_n} - f^*_b \right\|_{L_1} \\ &= \frac{h}{2N} \inf_{\tilde{b}_n \in \mathscr{C}} \max_{b \in \mathscr{C}} \mathbb{E}_b \left[ d_{\mathrm{H}}(\tilde{b}_n, b) \right]. \end{aligned}$$

Since $\{0,1\}^N_D$ is not the entire binary hypercube $\{0,1\}^N$, we cannot use Assouad's lemma. Instead, we will use Lemma 2 in conjunction with Birgé's bound. In order to do that, we must first construct a nice, well-separated subset of $\{0,1\}^N_D$. The following combinatorial result, due to P. Reynaud-Bouret, does the trick. There exists a subset $\mathscr{C} \subset \{0,1\}^N_D$ with the following properties:

1. $d_{\mathrm{H}}(b,b') > D/2$ for any two distinct $b, b' \in \mathscr{C}$;
2. $\log|\mathscr{C}| \geq \kappa D \log \frac{N}{D}$, where $\kappa \approx 0.233$.

Item 1 says that this $\mathscr{C}$ is $D/2$-separated in the Hamming distance. Thus, by Lemma 2,

$$\begin{aligned} R_n(h,\mathscr{F}) &\geq \frac{hD}{4N} \inf_{\tilde{b}_n \in \mathscr{C}} \max_{b \in \mathscr{C}} \mathbb{P}_b\left(\tilde{b}_n \neq b\right) \\ &= \frac{hD}{4N} \left(1 - \sup_{\tilde{b}_n \in \mathscr{C}} \min_{b \in \mathscr{C}} \mathbb{P}_b\left(\tilde{b}_n \neq b\right)\right). \end{aligned}$$

We are now in a position to apply Birgé's lemma:

$$\sup_{\tilde{b}_n \in \mathscr{C}} \min_{b \in \mathscr{C}} \mathbb{P}_b\left(\tilde{b}_n \neq b\right) \leq \alpha,$$

where $\alpha = 0.71$, provided

$$\bar{K} = \frac{1}{|\mathscr{C}| - 1} \sum_{b \in \mathscr{C},\, b \neq b_0} D\left(P^n_b \| P^n_{b_0}\right) \leq \alpha \log|\mathscr{C}|, \tag{40}$$

where $b_0$ is some fixed but arbitrary element of $\mathscr{C}$, and $P^n_b$ is a product of $n$ copies of $P_b$ (recall that our estimators are based on an i.i.d. sample of size $n$). So now we turn to the analysis of $\bar{K}$. For any two $b, b' \in \{0,1\}^N$, we have

$$\begin{aligned} D\left(P^n_b \| P^n_{b'}\right) &= nD(P_b \| P_{b'}) \\ &= n \sum_{i=1}^N \sum_{y \in \{0,1\}} P_b(x_i,y) \log \frac{P_b(x_i,y)}{P_{b'}(x_i,y)} \\ &= \frac{n}{N} \sum_{i=1}^N \sum_{y \in \{0,1\}} P^{(b)}_{Y|X}(y|x_i) \log \frac{P^{(b)}_{Y|X}(y|x_i)}{P^{(b')}_{Y|X}(y|x_i)} \\ &= \frac{n}{N} \sum_{i=1}^N D\left(\mathrm{Bernoulli}\left(\frac{1+(2b_i-1)h}{2}\right) \middle\| \mathrm{Bernoulli}\left(\frac{1+(2b'_i-1)h}{2}\right)\right), \end{aligned}$$

where the first line is by (27), while the rest follows from definitions. Now, in the above summation, the $i$th term is nonzero if and only if $b_i = b'_i$, and in that case it is equal to

$$D\left(\text{Bernoulli}\left(\frac{1+h}{2}\right)\middle\|\text{Bernoulli}\left(\frac{1-h}{2}\right)\right) = h\log\frac{1+h}{1-h}.$$

Consequently,

$$\begin{aligned} D\big(P_b^n\|P_{b'}^n\big) &= \frac{n}{N}h\log\left(\frac{1+h}{1-h}\right)\sum_{i=1}^{N}1_{\{b_i\neq b'_i\}} \\ &= \frac{n}{N}h\log\left(\frac{1+h}{1-h}\right)\cdot d_{\mathrm{H}}(b,b'). \end{aligned}$$

Now, for any two $b, b' \in \{0,1\}_D^N$, $d_{\mathrm{H}}(b,b') \le \mathrm{wt}(b)+\mathrm{wt}(b') = 2D$. Moreover, using the bound $\log t \le t-1$ for any $t \ge 0$, we have

$$h\log\frac{1+h}{1-h} \le h\left(\frac{1+h}{1-h}-1\right) = \frac{2h^2}{1-h}.$$

Therefore, for any two $b, b' \in \mathscr{C}$,

$$D\big(P_b^n\|P_{b'}^n\big) \le \frac{4nh^2D}{N(1-h)},$$

which implies that $\bar{K} \le \frac{4nh^2D}{N(1-h)}$. Using this and the fact that $\log|\mathscr{C}| \le \kappa D\log\frac{N}{D}$ with $\kappa \approx 0.233$, we see that the condition (40) will be satisfied if we can choose some $N \ge 4D$, so that

$$N\log\frac{N}{D} \ge \frac{4nh^2}{\alpha\kappa(1-h)}. \tag{41}$$

A tedious calculation (see [MN06]) shows that the choice

$$N = \left\lfloor \frac{8nh^2}{\alpha\kappa(1-h)\left(1+\log\frac{nh^2}{D}\right)} \right\rfloor$$

does the job, provided $h \ge \sqrt{D/n}$. This completes the proof.

## References


[Bir05] L. Birgé. A new lower bound for multiple hypothesis testing. *IEEE Transactions on Information Theory*, 51(4):1611–1615, April 2005.

[CT06] T. M. Cover and J. T. Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006.

[MN06] P. Massart and É. Nédélec. Risk bounds for statistical learning. *Annals of Statistics*, 34(5):2326–2366, 2006.

[Vap98] V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[Yu97] B. Yu. Assouad, Fano, and Le Cam. In D. Pollard, E. Torgersen, and G. Yang, editors, *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.